HHS Public Access

Author manuscript

Prev Sci. Author manuscript; available in PMC 2016 April 01.

Published in final edited form as:

Prev Sci. 2015 April ; 16(3): 475–485. doi:10.1007/s11121-014-0513-z.

Assessing the Generalizability of Randomized Trial Results to Target Populations

Elizabeth A. Stuart,
Johns Hopkins Bloomberg School of Public Health

Catherine P. Bradshaw, and
University of Virginia

Philip J. Leaf
Johns Hopkins Bloomberg School of Public Health

Abstract

Recent years have seen increasing interest in and attention to evidence-based practices, where the "evidence" generally comes from well-conducted randomized trials. However, while those trials yield accurate estimates of the effect of the intervention for the participants in the trial (known as "internal validity"), they do not always yield relevant information about the effects in a particular target population (known as "external validity"). This may be due to a lack of specification of a target population when designing the trial, difficulties recruiting a sample that is representative of a pre-specified target population, or to interest in considering a target population somewhat different from the population directly targeted by the trial. This paper first provides an overview of existing design and analysis methods for assessing and enhancing the ability of a randomized trial to estimate treatment effects in a target population. It then provides a case study using one particular method, which weights the subjects in a randomized trial to match the population on a set of observed characteristics. The case study uses data from a randomized trial of School-wide Positive Behavioral Interventions and Supports (PBIS); our interest is in generalizing the results to the state of Maryland. In the case of PBIS, after weighting, estimated effects in the target population were similar to those observed in the randomized trial. The paper illustrates that statistical methods can be used to assess and enhance the external validity of randomized trials, making the results more applicable to policy and clinical questions. However, there are also many open research questions; future research should focus on questions of treatment effect heterogeneity and further developing these methods for enhancing external validity. Researchers should think carefully about the external validity of randomized trials and be cautious about extrapolating results to specific populations unless they are confident of the similarity between the trial sample and that target population.

The authors declare they have no conflict of interest.

Clinical Trial Registry Number: NCT01583127

Keywords

Propensity scores; Horvitz-Thompson weighting; Positive Behavioral Interventions and Supports (PBIS)

Recent years have seen increasing interest in and attention to evidence-based practices, where the "evidence" generally comes from well-conducted randomized controlled trials (RCTs). However, while those trials yield accurate estimates of the effect of the intervention for the people in the trial, they do not always yield relevant information about the effects in a target population, such as the population relevant for a particular policy decision. This may be due to a lack of specification of a target population when designing the trial, difficulties recruiting a sample that is representative of a pre-specified target population, or to interest in considering a target population somewhat different from the population directly targeted by the trial. Standard methods and designs generally do not easily allow for the generalization of treatment effects from an RCT sample to a target population different from the population directly represented by the trial. This trade-off is the traditional distinction of "internal validity" versus "external validity" (Shadish, Cook, & Campbell, 2002). There are of course many reasons that effects seen in trials may not carry over to other target populations, including implementation difficulties, measurement differences, or different settings (Shadish, Cook, & Campbell, 2002). However, answering questions about generalizability and external validity is crucial for making trial results applicable more broadly, for both clinical and policy questions.

Although researchers and policymakers have made important advances in our understanding of the processes for disseminating and implementing programs, less attention has been paid to estimating what the effects of the programs will be in populations somewhat different from those participating in the original RCTs of a particular program. Individuals or organizations that volunteer to participate in research may differ in a number of ways from those targeted through wide-scale implementations of "effective practices." With a focus on prevention research, Flay et al. (2005) highlight the need to assess generalizability to establish evidence of effectiveness, but also acknowledge "The problem of generalizability remains an important area for scientific definition and investigation" (p. 164). In the current paper we discuss the conceptual issues associated with assessing the generalizability of RCT results, focusing on issues of external validity that may arise due to differences in the types of subjects participating in a trial and those in the target population.

More specifically, this paper discusses methods and diagnostics for investigating whether the effects seen in a trial would carry over to a population that may be somewhat different from the trial participants. This is a particularly important question for policymakers determining whether to recommend broad implementation of a particular program, and is especially relevant to prevention science. The methods also allow researchers and policy-makers to take advantage of existing RCTs and population data in order to predict impacts in those populations. The motivating example, discussed further below, estimates what the effects would be if all schools in the state of Maryland implemented the school-wide behavior improvement program Positive Behavioral Interventions and Supports (PBIS). This

question is related to, but not the same as, questions of heterogeneous treatment effects (links between these issues will be discussed further below). It is important to note, however, that our methods do not assume that program effects are constant across the population. So while the overall question is whether a program should receive broad implementation, a positive result would not necessarily imply that the program will equally benefit everyone in the population. In this way it is also a somewhat different question from a desire to tailor interventions based on individual characteristics. These are important issues that require more research attention, but are beyond the focus of the current review, which aims to answer questions about broad implementation.

Issues Associated with Representativeness in Randomized Trials

There has been increasing attention paid to the fact that people in an RCT may differ from those who eventually get the program or treatment of interest, especially in areas related to mental health (e.g., Humphreys et al., 2007; Rothwell, 2005; Stirman et al., 2005; Westen, 2006) and as highlighted in a recent *Nature* editorial (Nature, 2010). Braslow et al. (2005) found that few studies of psychiatric treatment enrolled representative samples and that in particular, minorities are often under-represented, and most studies do not address this limitation or even mention the representativeness of their samples in reports. Similarly, Wisniewski et al. (2009) found large differences between individuals enrolled in the STAR*D effectiveness trial and those who would have conceivably been enrolled in a more limited efficacy trial. However, there has been limited investigation into what to do about those differences, how to identify them, or how much they matter.

In addition, the topic of external validity has not been discussed as much in the social and behavioral sciences (with a few recent exceptions, including Olsen et al., 2013; O'Muircheartaigh & Hedges, 2014; Tipton, 2013). Some of the considerations may be quite different in different fields. For example, in medical contexts there are often many trials on the same topic, making research synthesis methods such as meta-analysis much more feasible than in the social and behavioral sciences, in which there are often only one or two trials of a particular program. Another distinction is that RCTs in the social and behavioral sciences (such as evaluations of educational interventions) often lack explicit inclusion and exclusion criteria, which may be due at least in part to a stronger focus on prevention–often universal prevention–programs. Generalizing results to subjects explicitly excluded from a trial requires specific methods (e.g., Pressler & Kaizar, 2013), and is not a topic we discuss further here.

Existing Methods for Assessing or Enhancing Generalizability

Existing methods for assessing or facilitating generalizability can be classified into two types: those related to the design of the trial, and after-the-fact analysis of the trial data.

Existing Study Design Strategies

The best way to ensure the generalizability of randomized trial results is to enroll a representative sample of subjects in the trial (Braslow et al., 2005). However, drawing a representative sample of subjects requires 1) knowing the population of interest in advance,

and 2) having a listing of, and access to, everyone in that population (and ideally with some characteristics observed on everyone in that population). Even when those criteria are met, random sampling from the population can be quite expensive, and, of course, works best when the selected subjects actually consent to participation. When selected subjects can decline to participate the benefits of having selected those subjects randomly may be lost (Shadish, 1995). In particular, the subjects that consent to be in an RCT may be quite different from the types of subjects that would implement a program once it is in general circulation, with some evidence of efficacy from that trial.

Random selection is thus most commonly used in large national evaluations of programs, where those programs are implemented in program sites. The program sites can then be selected randomly (although often with unequal probabilities of selection), and individuals within the selected sites randomized to treatment or control groups. This type of design is relatively rare (Olsen et al., 2013), but has been used in evaluations of Upward Bound (U.S. Department of Education, 2009), Head Start (U.S. Department of Health and Human Services, 2010), and Job Corps (Schochet, Burghardt, & McConnell, 2008).

Recent movement towards effectiveness trials carried out in real-world settings (Flay et al., 2005) offers a step toward generalizability. Effectiveness trials often enroll a broader range of subjects than do more narrow efficacy trials, and are typically conducted in a broader range of settings. However, although effectiveness trials are likely more representative of a population of interest than are efficacy trials, there is still no guarantee that the effectiveness trial will enroll subjects representative of some target population of interest. Practical clinical trials, which are large-scale randomized trials that aim to enroll a representative population (e.g., Insel, 2006), are a further step toward generalizability, but those trials require considerable time and money.

A final design strategy for enrolling representative subjects is “purposive sampling.” Shadish, Cook, and Campbell (2002) describe two types: “heterogeneous” and “typical.” Heterogeneous purposive sampling aims to enroll a set of heterogeneous subjects, to reflect the range of units that are in the target population. In contrast, “typical” purposive sampling aims to enroll subjects who are typical (or “average”) in the population. However, most of the work describing purposive sampling has been fairly conceptual, without much guidance on actually carrying it out, and when used it tends to be in a relatively informal way. Tipton et al. (2014) provide one potential approach for formalizing the idea of purposive sampling, using a multivariate distance measure to ensure that a study sample looks similar to the target population on a set of observed characteristics. However, these design strategies are not sufficient if the target population is different from the target population of the original trial; the trial may have been designed for one particular purpose, but future decision makers may be interested in using the results from the trial to inform decisions regarding other target populations. The analysis methods described further below can be used even in that case.

Existing Study Analysis Strategies

There are also a number of analytic strategies that have been developed to assess population treatment effects after a trial (or more commonly, a set of trials) has been carried out. Some

of these strategies consist simply of documenting various factors related to generalizability. Others, including post-stratification, meta-analysis-related methods, and a reweighting approach (the method we will focus on in our motivating example below), actually aim to estimate population treatment effects.

One broadly known strategy for documenting factors that may affect the generalizability of study results is the "Reach Effectiveness Adoption Implementation Maintenance" (Re-AIM) framework by Green and Glasgow (2006). Re-AIM provides simple summary measures of external validity, covering a wide range of areas (such as the percentage of individuals who participate, measures of attrition, quality of implementation) but with relatively little focus on the particular issue considered in this paper—that of differences in the characteristics of individuals in a trial sample and a target population.

Post-stratification is arguably the most straightforward and most common way of estimating population-level treatment effects from a randomized trial. Post-stratification estimates subgroup-specific treatment effects in the trial and then averages them using population weights to generate an average effect across the population; Rosenbaum (1987) refers to this as "direct adjustment." Post-stratification is commonly used in sample surveys as a way to calibrate the survey sample to population distributions (Holt & Smith, 1979); it has also been extended to estimating population-level effects. As a simple example, imagine the trial was 20% female and 80% male but the population of interest was 50/50. Post-stratification would take an equally weighted average of the male- and female-specific effect estimates from the trial as an estimate of what the effect would be in the population. The method is very straightforward but is limited in the number of variables for which adjustments can be made. For example, post-stratification cells can get very small even when trying to post-stratify on basic demographics such as gender, race, ethnicity, and age groups. Tipton (2013) presents a way of combining post-stratification with propensity scores in order to adjust for a larger set of variables; that approach is highly related to the weighting-based methods discussed further below.

Another set of methods is useful when there are multiple studies available on a particular topic. These include meta-analysis, research synthesis approaches, and the confidence profile method. The main purpose of a meta-analysis is to combine results from multiple studies, often used in settings where multiple randomized trials have been done on roughly the same research question and population (Hedges & Olkin, 1985; Sutton & Higgins, 2008). Meta-analysis typically takes just one result from each study, for example the treatment effect size, and combines those results across studies, either assuming that the effect sizes from the different studies are all estimating some common effect (a fixed effects model), or allowing variation in the effects across studies (a random effects model). A drawback of traditional meta-analysis is that it does not necessarily allow generalization of results to a target population, especially if the set of trials were all conducted on similar types of subjects, who may not be representative of the population. Traditional meta-analysis puts relatively little focus on assessing how similar the individuals in the trials are to the target population.

A broader class of methods, called research synthesis (or cross-design synthesis) involves similar ideas as meta-analysis, but is potentially more able to explicitly address the question of generalizability. Research synthesis enables the combination of results from randomized and non-randomized studies (Pressler & Kaizar, 2013; Prevost, Abrams, & Jones, 2000), and thus, for example, could combine information on program effects from a randomized trial with information from an observational study, which may contain a more representative sample (Imai et al., 2008). Research synthesis does this by modeling multiple parameters from each study and incorporating study characteristics into the model (e.g., Brown, Wang, & Sandler, 2008), including, for example, beliefs regarding the relative merits of the multiple sources of evidence (e.g., Turner et al., 2009). However, more work need is needed to fully investigate the potential use of research synthesis to answer the question of generalizability of interest in this paper, as estimating population effects are not always the explicit goal of these methods. For example, research synthesis approaches do not have formal ways of examining how similar subjects in a trial (or trials) are to individuals in a target population.

Some recent statistical approaches have more explicitly considered the question of how to estimate population-level treatment effects from randomized trial data. The most common strategies used are weighting approaches, which weight the trial subjects to resemble the population (e.g., Cole & Stuart, 2010; Stuart et al., 2011). Other researchers have examined similar approaches of weighting a sample to a population but only for the purpose of estimating population means, not causal effects (e.g., Pan & Schaubel, 2009). In the current paper, we propose that propensity scores (Rosenbaum & Rubin, 1983) can be used to summarize and assess the similarity of a randomized trial sample and a target population. This analytic approach provides a summary measure of the differences in the observed characteristics of the individuals in the trial and in the population. Then propensity score adjustment methods can be used to, first, examine how well those characteristics capture the differences between the sample and population, and then to estimate what the effects would be in the population. Below we apply this methodology to a group randomized controlled trial of a commonly used school-based prevention program, Positive Behavioral Interventions and Supports (PBIS)

Case Study: School-Wide Positive Behavioral Interventions and Supports

This work is motivated by an RCT of Positive Behavioral Interventions and Supports (PBIS), a school-wide non-curricular prevention framework that aims to improve school climate by creating improved systems and procedures that promote positive change in staff and student behaviors (Sugai et al., 2001). PBIS follows a 3-tiered public health approach to prevention, with a universal (school wide PBIS; SWPBIS) framework and more targeted (selective) and intensive (indicated) programs for students with higher need. To date most schools have focused on implementing the universal components of SWPBIS. Almost 20,000 schools are currently implementing SWPBIS, and the National Technical Assistance center on SWPBIS is funded by the U.S. Department of Education Office of Special Education Programs (www.pbis.org). However, there have been relatively few studies of the effectiveness of SWPBIS, including just two randomized trials (Bradshaw et al., 2009; Horner et al., 2009).

The current paper focuses on data from one of these trials, which was a group randomized effectiveness trial of SWPBIS (Murray, 1998), in which 37 Maryland elementary schools were randomized to receive SWPBIS or to be in a control condition (Bradshaw et al., 2009). Fidelity assessments indicated that schools in the PBIS condition in the trial implemented SWPBIS with high fidelity (Bradshaw et al., 2009). Follow-up of the participating schools over four years showed significant beneficial effects of PBIS on child behavior problems, including a reduction in office discipline referrals (ODRs), the outcome focused on in the current paper.

While there is interest in the immediate trial results, a relevant question for general policy is how effective SWPBIS is (or would be) across the state of Maryland (or the nation), not just in the 37 schools in the trial. Only public elementary schools in a limited number of school districts were eligible to participate in the trial, and all schools approached participated (Bradshaw, Waasdorp, & Leaf, 2012). However, observed factors that might limit the generalizablity of the trial results include the fact that the 37 trial schools had more children eligible for free or reduced price meals, less funding per pupil, and lower mathematics test scores as compared to all elementary schools in the state (Stuart et al., 2011). There is also some evidence of treatment effect heterogeneity in the trial, although primarily with respect to student-level characteristics such as students' grade when SWPBIS was introduced in the school (Waasdorp et al., 2012).

Formal Setting and Analytic Expression for Bias

We first provide some notation to facilitate our discussion. We consider a setting where a randomized trial has been conducted to estimate the effect of a program, P, relative to a control condition on a sample of subjects. Let S be an indicator variable for whether a subject is in the trial sample. By "program" we mean any intervention of interest, whether preventive or a treatment for a particular disorder or disease. The subjects in the sample may be individuals or they may be at a higher level, such as communities or schools, as in the case of SWPBIS. In the trial the program has been randomly assigned to subjects in the sample, forming a program group and a control group that are only randomly different from each other on all background characteristics. Interest is in determining the effectiveness of the program in a target population, represented by Ω, where the set with S=1 is a subset of Ω. In the SWPBIS example, the sample consists of the 37 schools in the effectiveness trial; Ω consists of the 717 elementary schools in the state. We assume that for all subjects in Ω (or a representative sample of them) we observe a set of background characteristics X, which describe both the subjects themselves and their broader contexts.

Ultimate interest is in the effect of the program P in the population Ω on an outcome Y. For each subject *i* in Ω, two potential outcomes exist. One is the value of the outcome if subject *i* receives the program, denoted $Y_i(1)$. The other is the value of the outcome if subject *i* receives the control condition, denoted $Y_i(0)$. The effect of the program for subject *i* is defined as $\tau_i = Y_i(1) - Y_i(0)$. For each subject *i* we can never observe both potential outcomes. For subjects that receive the program we observe $Y_i(1)$ and $Y_i(0)$ is known as the "counterfactual." The opposite holds for subjects that receive the control. While we thus cannot estimate subject-specific effects, it is possible to estimate the average effect for

groups of subjects. Averaging across all subjects in the population, the average effect of the program is $\tau_\Omega = \frac{1}{N_\Omega}\sum_{i=1}^{N_\Omega}(Y_i(1) - Y_i(0))$, where N_Ω is the number of subjects in the population. Imai, King, and Stuart (2008) call this the "Population Average Treatment Effect" (PATE) and this is our quantity of interest. Unfortunately, however, we cannot directly estimate the PATE, and in fact we may not observe Y_i (1) or Y_i (0) for many (or all) subjects in Ω. In the trial we observe Y_i (1) for the subjects in the program group and Y_i (0) for the subjects in the control group. From this data in the trial we can obtain an unbiased estimate of the average effect of the program in the sample: $\tau_S = \frac{1}{N_S}\sum_{i=1}^{N_\Omega}(Y_i(1) - Y_i(0)|S_i = 1)$, where N_s is the number of subjects in the sample. This quantity is sometimes termed the "Sample Average Treatment Effect" (SATE).

The difference between SATE and PATE can give us an expression for the bias in using SATE as an estimate of PATE. Olsen et al. (2013) present a conceptual model for participation in randomized experiments and show that the external validity bias can be expressed as a function of three population quantities: $Bias_X = \sigma_\Delta cv_P \rho_{\Delta P}$, where σ_Δ is a measure of the variability in treatment effects (in particular, their standard error) across the population, cv_P is a measure of the variability in the probabilities of being in the sample (specifically, the coefficient of variation), and $\rho_{\Delta P}$ is the correlation between the participation probabilities and treatment effects. The external validity bias is thus 0 if treatment effects do not vary (i.e., if there is no treatment effect heterogeneity), if the probabilities of participation do not vary, or if the probability of being in the sample is unrelated to treatment effect size. However, the bias will increase with more effect heterogeneity, more variability in the probability of participating, and higher correlations between those factors.

Under a particular form of the outcome model (a linear model) we can generate a more specific expression for the bias (Cole & Stuart, 2010). Assume a simple setting with outcome Y, treatment indicator T, binary covariate Z, and indicator S of being in the trial sample, where the treatment effect varies across levels of Z. In particular, let us assume an outcome model of the form $E(Y_i) = b_0 + b_T T + b_Z Z + b_{TZ} TZ$. The parameter b_{TZ} reflects the amount of treatment effect heterogeneity, with b_T the treatment effect for individuals with Z=0 and $b_T + b_{TZ}$ the treatment effect for individuals with Z=1. When $b_{TZ}=0$ there is no effect heterogeneity and thus no bias when using the SATE as an estimate of the PATE. Cole and Stuart (2010) derive an expression for the external validity bias when using SATE as an estimate of PATE (i.e., when using the simple difference in means of the outcome between treatment and control groups in the trial sample [the SATE] as an estimate of the average difference in treatment and control potential outcomes in the population [the PATE]). They show that this bias can be written as $b_{TZ} * \frac{P(Z=1)}{P(S=1)} * (P(S=1|Z=1) - P(S=1))$. This expression shows that the bias is a function of 1) the extent of the effect heterogeneity across levels of Z, as measured by b_{TZ}, 2) the proportion of individuals in the sample (P(S=1)), 3) the prevalence of the heterogeneity characteristic Z in the population (P(Z=1)), and 4) the degree to which

participation in the trial is associated with Z, as measured by P(S=1|Z=1)-P(S=1). The bias is 0 if 1) the entire population is in the trial sample (P(S=1)=1), 2) there is no heterogeneity in Z (P(Z=1)=0), or 3) Z is unrelated to participation in the trial (P(S=1|Z=1)=P(S=1)). These latter two factors have obvious analogs in the Olsen et al. (2013) bias expression provided above.

Figure 1 examines this bias formula across a few settings, illustrating how the bias increases as a function of these parameters. In particular, the three scenarios in Figure 1 show the consequences of different values of P(S), the proportion of the sample that participates, P(Z), the prevalence of the heterogeneity parameter in the population, and P(S|Z), the relationship between the heterogeneity characteristic Z and participation, across a range of values of b_{TZ}. The three panels vary in P(S) and P(Z), and each explores three values for the odds ratio of the relationship between participation in the trial (S) and the heterogeneity characteristic (Z).

A few points are evident from Figure 1. First, comparing the y-axes of the three panels (which are all on the same scale), the bias is much lower when P(S) is large (.8) than when it is small (.1), which is expected—the potential for bias is generally smaller if a larger proportion of the population is in the randomized trial sample. (However, we also note that P(S)=.8 is highly implausible for most real-world studies). Second, more heterogeneous treatment effects (larger values of b_{TZ}) increase the bias, as indicated by the positive slopes on all of the lines. Third, when S and Z are not strongly related (i.e., when participation in the trial is only weakly related to the heterogeneity characteristic), the bias is relatively small (the bottom lines on each panel). The weighting methods we propose below will be used to make the trial sample and population more similar with respect to Z, essentially trying to lower the odds ratio between S and Z to 1, and thus improving population treatment effect estimates.

Using Propensity Score Methods to Assess Generalizability

We now illustrate a method researchers can use to begin assessing the generalizability of their RCTs and estimate population treatment effects. This approach utilizes *propensity scores*, which are generally used for estimating causal effects in non-experimental settings (Rosenbaum & Rubin, 1983). Propensity scores in particular are used here to compare the subjects and contexts in the trial sample with those in the target population. The propensity score is typically defined as the probability of receiving some program (or “treatment”) versus a comparison condition, given a set of observed baseline characteristics. These characteristics may include both individual and contextual level variables. Propensity score matching can help ensure that the program and comparison subjects in a non-randomized study are as similar as possible. This is done by comparing groups of subjects with similar propensity scores, who, by virtue of the properties of the propensity score, will also have similar distributions of the observed background covariates (Rosenbaum & Rubin, 1983). In this way, propensity scores attempt to replicate a randomized experiment by comparing subjects across treatment conditions who have no systematic differences on the observed background characteristics. As explored by Cole and Stuart (2010) and Stuart et al. (2011), propensity score methods can also be used to examine the similarity of subjects in an RCT

sample and in a target population. Here, in order to summarize differences between the sample and population, the propensity score models membership in the RCT sample rather than receipt of the treatment, as is more common.

Propensity Scores as a Diagnostic—The first way in which propensity scores can be used is as a diagnostic to assess generalizability. Stuart et al. (2011) describe two diagnostic measures. The first is the average difference in propensity scores between the sample and population, divided by the standard deviation of the propensity scores. Like a discriminant function, due to the properties of the propensity score this provides the maximum difference in a linear combination of the covariates, giving researchers a one-number summary measure of the differences between a sample and a population, rather than requiring an examination of each covariate one at a time. Previous research on propensity scores has established 0.1 or 0.2 standard deviations as a difference that indicates unacceptable differences between the groups (in this case, the sample and the population), which would result in extrapolation and model dependence (Rubin, 2001; Stuart, 2010).

To estimate this quantity in the SWPBIS example we first fit a logistic regression model predicting participation in the trial as a function of the set of school characteristics in Table 1. We then calculated the predicted probability of participating in the trial for each school in the state population. As our measure of similarity we take the difference in average predicted participation probabilities for those schools in the trial and those schools not in the trial divided by the standard deviation in the participation probabilities; this measure was 0.73, indicating that the trial and population schools were nearly three quarters of a standard deviation apart. However, the real question is how well the observed characteristics capture the relevant differences between the trial and population schools. To investigate that, Stuart et al. (2011) compared the outcomes observed in the population schools not implementing SWPBIS to a weighted average of the outcomes in the control group of the randomized trial (who also were not implementing SWPBIS). If the observed characteristics adequately capture the differences between trial and population schools, the weighted control group mean should be similar to that observed in the population. (The details on these weights are provided below). In the SWPBIS example Stuart et al. (2011) found that, although the schools in the trial appear quite different on 3rd and 5th grade outcomes when not weighted, when weighted, the trial schools in the control group reflect what was happening statewide, especially for the 3rd grade measures.

Propensity Score Weighting to Estimate the PATE—We now extend Stuart et al. (2011) to obtain an estimate of the PATE, using methods similar to those in Cole and Stuart (2010). This uses a similar approach to the diagnostic described previously, but where both groups (treatment and control) are weighted to the population. The method is related to Horvitz-Thompson weighting in sample surveys (1952), in which surveyed individuals are weighted by their inverse probabilities of selection, and Inverse Probability of Treatment Weighting (IPTW), which is used in non-experimental studies to make the treatment and control groups comparable. In this context where the goal is to account for participation in the trial we refer to it as Inverse Probability of Participation Weighting (IPPW; Cole & Stuart, 2010). The crucial assumption underlying this approach is that the weight model

includes all variables that are associated with participation in the trial and that moderate treatment effects (Cole & Stuart, 2010; Stuart et al., 2011). In other words, that the observed characteristics are sufficient for generalizing treatment effects from the sample to the population. We also rely on a model of participation in the trial. The implications of these assumptions are discussed further below.

To estimate the PATE, we use the following procedure: 1) fit a model predicting participation in the RCT sample as a function of baseline characteristics, 2) define weights $w_i=1/p_i$, where p_i is the probability that subject *i* is in the RCT, obtained from that fitted model, and 3) run a weighted regression model using the trial subjects and their weights w_i. Step 3 essentially involves running the same analysis as is done to estimate the SATE in the trial, but now using the weights. (Note that in usual IPTW contexts the control group receives a weight of $1/(1-p_i)$. In our context here, only the trial data is used for estimating the population treatment effect estimates (since that is where the treatment and outcome are observed), and so the only individuals in the analysis are those in the trial sample). The Appendix of Cole and Stuart (2010) provides a proof for the consistency of this approach for estimating the PATE.

We now apply this procedure to the SWPBIS example where we aim to estimate the population effect of SWPBIS on Office Disciplinary Referrals (ODRs). ODRs were measured using a question on the Teacher Observation of Classroom Adaptation-Checklist (TOCA-C; Koth, Bradshaw, & Leaf, 2009), a binary report of whether each student had received an ODR during that school year. This teacher-reported measure of ODRs has been shown to be a valid indicator, as compared with administrative data (Pas, Bradshaw, & Mitchell, 2011). The teacher-report measure was collapsed over the 5 time periods available to create a binary 0/1 variable for each student: ever vs. never received an ODR.

In particular, we first estimated a model of participation in the trial, using a dataset with one observation for every elementary school in the state of Maryland (*N*=717) and an indicator for the 37 schools in the SWPBIS trial. We then fit a logistic regression model of participation in the trial as a function of school characteristics (demographics and test scores), measured in 2002 (see Table 1). Each school in the trial was then given a weight of 1/p, where p was the predicted probability from that logistic regression: their predicted probability of being in the trial. Table 1 investigates how well those weights work to make the trial and population schools look similar, showing the population means and the unweighted and weighted standardized mean differences comparing trial and population schools for all variables used in the propensity score model. The standardized mean difference for each variable is the (weighted or unweighted) difference in means, divided by the standard deviation. Commonly used in the propensity score literature to summarize the similarity of treatment and comparison groups, it can be used as a guide for whether groups are sufficiently similar for comparison. Table 1 shows that, after weighting, all standardized mean differences are below the 0.2 cutoff frequently used as an indication of sufficient similarity in the propensity score literature (Stuart, 2010).

To estimate treatment effects we used these IPTW weights in a logistic regression model predicting ODRs as a function of treatment status (SWPBIS vs. not), student, and school

characteristics, fit among the schools in the trial. For this we use student-level data and account for the clustering of students within schools using a multilevel (random effects) model. Because schools were the unit of participation for the trial, and thus the weights are at the school level, we run the model with the weights set at the school level (Level 2). The model controlled for the same student-level (grade cohort, gender, ethnicity, special education status, eligible for Free or Reduced Price Meals) and school-level (student mobility, student/teacher ratio, enrollment, faculty turnover) characteristics as in Bradshaw et al. (2012). The weights were calculated in the R software package (R Core Team, 2013) and the multilevel models were run using the gllamm procedure (Rabe-Hesketh, Skrondal, & Pickles, 2004) for Stata 12 (StataCorp, 2011).

In an unweighted model that estimates the SATE, the estimated effect of PBIS on ODRs was an odds ratio of 0.64 (95% CI: .55, 75; p=.00). (The SATE estimate is slightly different from that in Bradshaw et al. (2012, Table 6) because of slight estimation differences; Bradshaw et al. (2012) fit models in HLM (Raudenbush et al., 2011), while we used gllamm because we had easier access to gllamm. For reference, Bradshaw et al. (2012) reported an odds ratio of 0.66 ($p < .01$). The weighted estimate, estimating the PATE, was similar: 0.61 (95% CI: .53, .71; p=.00). The PATE point estimate is just slightly attenuated, but both the SATE and PATE indicate a reduction in ODRs for students in schools implementing SWPBIS. The similarity in odds ratios is likely in part a function of the lack of evidence of heterogeneity in treatment effects of the SWPBIS intervention on ODRs (Bradshaw et al., 2012). Using the weights allows us to estimate what the effects of SWPBIS would be if implemented statewide, accounting for the differences in observed characteristics between the schools in the trial and those statewide.

Conclusions

This paper presents one of the first formal discussions of external validity in the prevention science literature. As studies aiming to estimate causal effects become more and more rigorous in terms of their internal validity, and sometimes with respect to external validity for the original target population, we can now start thinking about ways to generalize those trial results to more formal, or slightly different, target populations. In particular, as interest increases in dissemination and implementation of prevention and treatment programs, it is more important than ever to understand how well existing studies might inform policy decisions in other contexts and locations, among varying populations. Other applications include contexts where the types of individuals in the target population change, for example, as more individuals become eligible for health insurance, thus changing the types of individuals seeking services.

A limitation of the existing work is that it can only account for differences between a sample and population in characteristics that are observed in the trial sample and the target population. In the SWPBIS trial many characteristics are observed on the schools in the trial (e.g., school climate), but in our statistical adjustments we were limited to only those characteristics also observed on all schools across the state, which consists primarily of demographics and test scores. In this example we found that the estimated population effects are similar in magnitude to the impacts estimated in the RCT. This may in part be because

there is limited evidence of treatment effect heterogeneity in the trial; specifically, Table 6 of Bradshaw et al. (2012) shows no variation in effects on ODRs across grade cohorts or special education status. Although there was some evidence of variation by gender, gender was a variable on which there was little variation across schools, and thus the trial and state schools are already well matched on that factor. The theory of the intervention leads us to believe that we likely measured the characteristics that may moderate treatment effects (e.g., school size, percent of students eligible for Title 1). However, other potentially important variables (such as principal's support for the program) are not observed on schools statewide. In addition, although evidence from the two existing RCTs provides information about the overall effectiveness of SWPBIS, little is known about whether treatment effects vary as a function of school characteristics, in part due to limited statistical power. Future work should further investigate whether program effects vary (and across which factors, both individual-level and contextual), and account for those factors when assessing external validity. The current paper also only considers pre-treatment factors that may moderate effects; additional work should consider how changes in factors such as take-up or participation rates might affect population treatment effects (e.g., Frangakis, 2009). Critical for assessing generalizability to a target population is identifying that population and having data on that population. In the SWPBIS example we had access to data on the population of Maryland schools. In other cases such extensive data is not available. However, methods such as those used in Cole and Stuart (2010) can be used to generate a pseudo-population using just the cross-tabulation of a few key covariates; in Cole and Stuart (2010) the population data available was simply the age by sex by race distribution in the target population. This also highlights the need for high quality population-level datasets that can be used to characterize potential target populations.

Methods for assessing and enhancing external validity are just beginning to be developed and thus, there are many directions for future research. In particular, much of the discussion around external validity rests on understanding treatment effect heterogeneity and what factors moderate treatment effects. If there is no effect heterogeneity the sample treatment effect can be assumed to generalize to the population (since then effects are constant across everyone); however, existing trials provide relatively little information regarding the extent of treatment effect heterogeneity. Especially given the limited power in most RCTs for detecting effect modification, new statistical methods are needed to help identify whether effects vary, and over what factors. Another question for future research is which characteristics to prioritize when equating the sample and the target population; the propensity score weighting used here prioritizes characteristics by how predictive characteristics are of participation in the trial (through the logistic regression of participation in the trial as a function of the characteristics). Another possibility would be to prioritize variables by how predictive they are of treatment effects, or of outcomes (e.g., using prognostic scores; Hansen, 2008). Finally, especially when the target population is very large and diverse (e.g., from the Census), or if the sample and population differ substantially, there is the potential for very large weights, which can cause poor performance of the weighting. This issue has been investigated in the context of propensity score weighting in non-experimental studies (Kang & Schafer, 2007), and should be examined in the generalizability context as well, where it may be an even larger concern. However, some

of the lessons from the standard IPTW context (such as weight trimming or standardization) may be able to be carried over to this new setting. These represent important directions for further statistical research.

RCTs provide unbiased estimates of the treatment effect in the sample at hand. However, often decision makers and policy makers are actually interested in what the treatment effect would be in some other, target, population. Existing methods and study designs are somewhat limited in their ability to estimate the effects in a target population. We hope that this paper prompts more discussion of these issues and helps researchers pay more attention to the issue of generalizing treatment effects from one sample to a target population.

Acknowledgments

Funding Source: Support for this project comes from grants from the Centers for Disease Control and Prevention (R49/CCR318627, 1U49CE 000728, and K01CE001333), the National Institute of Mental Health (1R01MH67948; K25 MH083846), and the Institute of Education Sciences (R305A090307).

References

Bradshaw CP, Koth CW, Thornton LA, Leaf PJ. Altering school climate through school-wide positive behavioral interventions and supports: Findings from a group-randomized effectiveness trial. Prevention Science. 2009; 10(2):100–115. [PubMed: 19011963]

Bradshaw CP, Waasdorp TE, Leaf PJ. Effects of school-wide positive behavioral interventions and supports on child behavior problems. Pediatrics. 2012; 130(5):1136–1145. [PubMed: 23129082]

Braslow JT, Duan N, Starks SL, Polo A, Bromley E, Wells KB. Generalizability of studies on mental health treatment and outcomes, 1981–1996. Psychiatric Services. 2005; 56(10):1261–1268. [PubMed: 16215192]

Brown CH, Wang W, Sandler I. Examining how context changes intervention impact: The use of effect sizes in multilevel mixture meta-analysis. Child Development Perspectives. 2008; 2(3):198–205. [PubMed: 20585469]

Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: The ACTG-320 trial. American Journal of Epidemiology. 2010; 172:107–115. [PubMed: 20547574]

Flay BR, Biglan A, Boruch RF, Castro FG, Gottfredson D, Kellam S, Mościcki EK, Schinke S, Valentine JC, Ji P. Standards of evidence: Criteria for efficacy, effectiveness, and dissemination. Prevention Science. 2005; 6(3):151–175. [PubMed: 16365954]

Frangakis CE. The calibration of treatment effects from clinical trials to target populations. Clinical Trials. 2009; 6:136–140. [PubMed: 19342466]

Green LW, Glasgow RE. Evaluating the relevance, generalization, and applicability of research: Issues in external validation and translation methodology. Evaluation & the Health Professions. 2006; 29:126–153. [PubMed: 16510882]

Hansen BB. The prognostic analogue of the propensity score. Biometrika. 2008; 95:481–488.

Hedges, LV.; Olkin, I. Statistical methods for meta-analysis. Orlando, FL: Academic Press; 1985.

Holt D, Smith TMF. Post stratification. Journal of the Royal Statistical Society, Series A. 1979; 142(1):33–46.

Horner RH, Sugai G, Smolkowski K, Eber L, Nakasato J, Todd AW, Esperanza J. A randomized, wait-list controlled effectiveness trial assessing school-wide positive behavior support in elementary schools. Journal Positive Behavior Interventions. 2009; 11(3):133–144.

Horvitz D, Thompson D. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association. 1952; 47:663–685.

Humphreys K, Weingardt KR, Harris AHS. Influence of subject eligibility criteria on compliance with national institutes of health guidelines for inclusion of women, minorities, and children in treatment research. Alcoholism: Clinical and Experimental Research. 2007; 31(6):988–995.

Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. Journal of the Royal Statistical Society Series A. 2008; 171:481–502.

Insel TR. Beyond efficacy: The STAR*D trial. American Journal of Psychiatry. 2006; 163:5–7. [PubMed: 16390879]

Kang JD, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science. 2007; 22(4):523–539.

Koth CW, Bradshaw CP, Leaf PJ. Teacher observation of classroom adaptation-checklist: Development and factor structure. Measurement and Evaluation in Counseling and Development. 2009; 42(1):15–30.

Murray, DM. Design and analysis of group-randomized trials. New York: Oxford Press; 1998.

Nature. Editorial: Putting gender on the agenda. Nature. 2010; 465(7299):665.

Olsen R, Bell S, Orr L, Stuart EA. External validity in policy evaluations that choose sites purposively. Journal of Policy Analysis and Management. 2013; 32(1):107–121. [PubMed: 25152557]

O'Muircheartaigh C, Hedges LV. Generalizing from unrepresentative experiments: A stratified propensity score approach. Journal of the Royal Statistical Society, Series C, Applied Statistics. 2014 Early view online. 10.1111/rssc.12037

Pan Q, Schaubel DE. Evaluating bias correction in weighted proportional hazards regression. Lifetime Data Analysis. 2009; 15:120–146. [PubMed: 18958616]

Pas E, Bradshaw CP, Mitchell MM. Examining the validity of office discipline referrals as an indicator of student behavior problems. Psychology in the Schools. 2011; 48(6):541–555.

Pressler TR, Kaizar EE. The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. Statistics in Medicine. 201310.1002/sim.5802

Prevost TC, Abrams KR, Jones DR. Hierarchical models in generalized synthesis of evidence: An example based on studies of breast cancer screening. Statistics in Medicine. 2000; 19(24):3359–3376. [PubMed: 11122501]

R Core Team. R: A language and environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2013. Retrieved from the R project website: http://www.R-project.org

Rabe-Hesketh S, Skrondal A, Pickles A. Generalized multilevel structural equation modelling. Psychometrika. 2004; 69(2):167–190.

Raudenbush, SW.; Bryk, AS.; Cheong, YF.; Congdon, RT., Jr; du Toit, M. Hierarchical linear and nonlinear modeling (HLM7). Lincolnwood, IL: Scientific Software International, Inc; 2011.

Rosenbaum PR. Model-based direct adjustment. Journal of the American Statistical Association. 1987; 82(398):387–394.

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70(1):41–55.

Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply? Lancet. 2005; 365(9453):82–93. [PubMed: 15639683]

Rubin DB. Using propensity scores to help design observational studies: Application to the tobacco litigation. Health Services & Outcomes Research Methodology. 2001; 2:169–188.

Schochet PZ, Burghardt J, McConnell S. Does job corps work? Impact findings from the national job corps study. American Economic Review. 2008; 98(5):1864–86.

Shadish WR. The logic of generalization: Five principles common to experiments and ethnographies. American Journal of Community Psychology. 1995; 23(3):419–428.

Shadish, WR.; Cook, TD.; Campbell, DT. Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin Company; 2002.

StataCorp. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP; 2011.

Stirman SW, Derubeis RJ, Crits-Christoph P, Rothman A. Can the randomized controlled trial literature generalize to nonrandomized patients? Journal of Consulting and Clinical Psychology. 2005; 73(1):127–35. [PubMed: 15709839]

Stuart EA. Matching methods for causal inference: A review and a look forward. Statistical Science. 2010; 25(1):1–21. [PubMed: 20871802]

Stuart EA, Cole S, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. Journal of the Royal Statistical Society Series A. 2011; 174(2): 369–386.

Sugai, G.; Horner, R.; Gresham, F. Behaviorally effective school environments. In: Shinn, M.; Stoner, G.; Walker, H., editors. Interventions for academic and behavior problems: Preventive and remedial approaches. Silver Spring, MD: National Association of School Psychiatrists; 2001. p. 315-350.

Sutton AJ, Higgins JP. Recent developments in meta-analysis. Statistics in Medicine. 2008; 27(5):625–650. [PubMed: 17590884]

Tipton E. Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. Journal of Educational and Behavioral Statistics. 2013; 38(3):239–266.

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

Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. Journal of the Royal Statistical Society, Series A. 2009; 172(1):21–47.

U.S. Department of Education. The Impacts of Regular Upward Bound on Postsecondary Outcomes Seven to Nine Years After Scheduled High School Graduation. Washington, D.C: Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service; 2009.

U.S. Department of Health and Human Services. Head Start Impact Study Final Report. Washington, D.C: Office of Planning, Evaluation and Policy Development, Administration for Children and Families, Policy and Program Studies Service; 2010.

Waasdorp TE, Bradshaw CP, Leaf PJ. The impact of Schoolwide Positive Behavioral Interventions and Supports on bullying and peer rejection. Archives of Pediatric and Adolescent Medicine. 2012; 166(2):149–156.

Westen, DI.; Stirman, SW.; DeRubeis, RJ. Are Research Patients and Clinical Trials Representative of Clinical Practice?. In: Norcross, JC.; Beutler, LE.; Levant, RF., editors. Evidence-based practices in mental health: Debate and dialogue on the fundamental questions. Washington, DC: American Psychological Association; 2006. p. 161-189.

Wisniewski S, Rush A, Nierenberg A, Gaynes B, Warden D, Luther J, McGrath PJ, Lavori PW, Thase ME, Fava M, Trivedi MH. Can phase III trial results of antidepresseant medications be generalized to clinical practice? A STAR*D report. American Journal of Psychiatry. 2009; 166(5):599–607. [PubMed: 19339358]

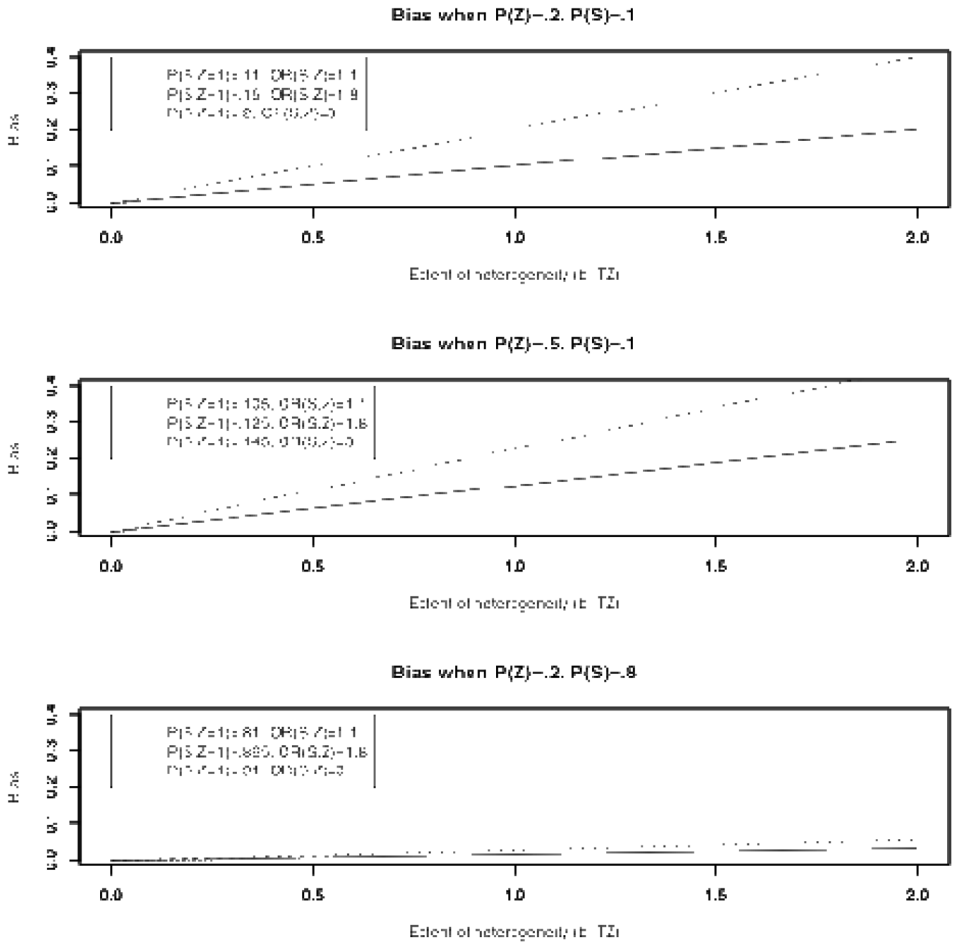

Figure 1.

Bias in estimating the Population Average Treatment Effect using a randomized trial sample when there is effect heterogeneity and the probability of participation in the trial is related to the characteristic driving treatment effects. P(S) denotes the probability of participation in the trial; P(Z) refers to the prevalence of a characteristic related to treatment effects and participation, P(S|Z=1) (and the related quantity reflecting the odds ratio between S and Z, OR(S,Z)) reflects the degree of association between the heterogeneity characteristic Z and participation S, and b_TZ reflects the degree of association between Z and treatment effects, as expressed in the form of the outcome model: $E(Y_i) = b_0 + b_T T + b_Z Z + b_{TZ} TZ$.

Table 1

Characteristics of schools in PBIS trial and those in state population

School characteristic	Mean in Trial	Mean in Population	Unweighted Standardized Difference in Means	Weighted Standardized Difference in Means
Attendance rate	95.3	95.3	0.02	0.06
Enrollment	485	480	0.03	0.07
% students eligible for Free or Reduced Price Meals	39.7	36.3	0.12	−0.07
% students in special education	13.8	15.1	−0.09	−0.13
% students eligible for Title 1	47.3	27.4	0.47	−0.02
% students White	60.3	54.1	0.18	0.11
3rd grade math	27.4	31.9	−0.22	−0.06
3rd grade reading	32.9	34.5	−0.08	0.02
5th grade math	44.6	51.0	−0.21	−0.11
5th grade reading	54.2	53.3	0.04	−0.03
% students suspended	6.3	4.6	0.33	0.18
N	37	717		

Note: Weighted means reflect IPTW weighting of schools in trial. Standardized difference in means refers to difference in means (trial minus population) divided by standard deviation. Test score variables reflect the % of students scoring Proficient or Advanced on the Maryland state standardized test. All variables measured in 2002, before the trial began.