Journal of the Royal Statistical Society
SERIES A Statistics in Society

J. R. Statist. Soc. A (2011)
174, *Part* 2, *pp.* 369–386

The use of propensity scores to assess the generalizability of results from randomized trials

Elizabeth A. Stuart,

Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

Stephen R. Cole,

University of North Carolina, Chapel Hill, USA

and Catherine P. Bradshaw and Philip J. Leaf

Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

[Received March 2010. Revised August 2010]

Summary. Randomized trials remain the most accepted design for estimating the effects of interventions, but they do not necessarily answer a question of primary interest: will the programme be effective in a target population in which it may be implemented? In other words, are the results generalizable? There has been very little statistical research on how to assess the generalizability, or 'external validity', of randomized trials. We propose the use of propensity-score-based metrics to quantify the similarity of the participants in a randomized trial and a target population. In this setting the propensity score model predicts participation in the randomized trial, given a set of covariates. The resulting propensity scores are used first to quantify the difference between the trial participants and the target population, and then to match, subclassify or weight the control group outcomes to the population, assessing how well the propensity-score-adjusted outcomes track the outcomes that are actually observed in the population. These metrics can serve as a first step in assessing the generalizability of results from randomized trials to target populations. The paper lays out these ideas, discusses the assumptions underlying the approach and illustrates the metrics by using data on the evaluation of a schoolwide prevention programme called 'Positive behavioral interventions and supports'.

Keywords: Causal inference; External validity; 'Positive behavioral interventions and supports'; Research synthesis

1. Introduction

Randomized trials remain the most accepted design for estimating the effects of interventions, and they serve as the basis for the recommendations being made for prevention and treatment programmes by the US Department of Education (the 'What works clearinghouse'), the Substance Abuse and Mental Health Services Administration (the 'National registry of evidence-based programs and practices'), the Agency for Healthcare Research and Quality (evidence-based practice centres) and researchers (the Cochrane Collaboration). Although crucial for assessing efficacy, randomized trials do not necessarily answer a question of primary policy interest: will the programme be effective in a target population in which it may be implemented? In other words, are the results generalizable? Efficacy trials assess whether an

Address for correspondence: Elizabeth A. Stuart, Departments of Mental Health and Biostatistics, Johns Hopkins Bloomberg School of Public Health, 8th Floor, 624 North Broadway, Baltimore, MD 21205, USA.
E-mail: estuart@jhsph.edu

 0964–1998/11/174369

intervention works under ideal circumstances, often including a rather homogeneous set of participants (Flay, 1986). As defined by Campbell (1957), page 297,

> 'The second criterion is that of external validity, representativeness, or generalizability: to what populations, settings, and variables can this effect be generalized?'.

Effectiveness trials are a step in the direction of generalizability, in that they assess whether the intervention works in real world conditions, with a broader set of participants (Flay, 1986). However, even effectiveness trials rarely are done by using participants who are representative of the target populations in which the interventions being evaluated may eventually be implemented (Rothwell, 2005a). Statistical methods to assess the generalizability of results from effectiveness trials to target populations are needed (National Institute of Mental Health, 1999; Institute of Medicine, 2006).

This paper focuses on the aspect of generalizability that is related to differences in the characteristics of participants in an effectiveness trial and in a target population. Participants in effectiveness trials are rarely representative of the target population of interest and effects often vary for different types of people and in different contexts. This combination means that the results that are seen in a randomized trial may not reflect the effects that would be seen if the intervention were implemented in a different target population (Flay *et al.*, 2005). As an example, one explanation for discrepancies regarding the effects of hormone replacement therapy for post-menopausal women in the 'Women's health initiative' randomized trial and the 'Nurse's health study' observational study is differences in the types of women in the two studies (Grodstein *et al.*, 2003), although other possible explanations have also been provided (Hernán *et al.*, 2008). Another example relates to recommendations concerning breast conservation *versus* mastectomy for women with breast cancer; the General Accountability Office was concerned that the results that are obtained from randomized trials may not carry over to women in general medical practice and so also used observational data methods to estimate the effects for a broader group of women (Rubin, 2008). This issue is similar to that of the representativeness of results from Web-based surveys where respondents opt in for participation; some recent work has investigated implications for survey research, but without a focus on studies estimating causal effects (Couper and Miller, 2008).

The work in this paper is related to proposals to use weighting-based approaches to estimate effects for a target population by using data from a randomized trial (e.g. Shadish *et al.* (2002), Cole and Stuart (2010), Haneuse *et al.* (2009) and Pan and Schaubel (2009)). These proposals use ideas that are similar to Horvitz–Thompson estimation for sample surveys and inverse probability of treatment weighting (IPTW) for non-experimental studies (Lunceford and Davidian, 2004), but where the trial sample is weighted to represent the population. In this paper we take a step back to develop diagnostics to help researchers to determine when such generalizations may be possible and reliable, and we also investigate other propensity score approaches in addition to weighting. Almost no metrics for this purpose are currently available. Glasgow *et al.* (2006) discussed the need for considering and measuring the 'reach' of an intervention in terms of participation and representativeness of patients but they discussed only very simple metrics. The state of the art currently is simply to compare the covariates one by one and to make qualitative statements regarding the similarity of subjects in a trial and some target population, but it can be difficult to summarize across many covariates. The work that is described here aims to provide summary measures of representativeness with respect to observed pretreatment characteristics. In particular, we investigate the use of propensity scores to measure and quantify differences between the participants in a randomized trial and a target population and we use results from the propensity score literature on how to quantify differences between two groups

and to determine how large a difference is too much for reliable comparison, applying those results to a new area.

The methods proposed are illustrated by using a randomized trial of a schoolwide behaviour improvement programme: 'Positive behavioral interventions and supports' (PBIS) (Sugai and Horner, 2006). The trial involved the random assignment of 37 public elementary schools from five Maryland school districts to the PBIS programme or a control condition (Bradshaw *et al.*, 2009). Primary outcomes of interest include behaviour and academic performance, as measured by student discipline problems, school climate and student test scores. We also take advantage of school level data that are available for all elementary schools across the state of Maryland. The question of interest is how similar the schools in the trial are to those across the state, and whether the results from the trial might hold across the state of Maryland; this is the policy question that is of interest to policy makers deciding whether or not the PBIS programme should be recommended or implemented state wide. This paper focuses on the statistical ideas and concepts; future work will discuss more of the substantive issues that are associated with the PBIS programme itself and generalizing its effects.

This paper outlines the main idea behind using propensity scores to measure similarity of participants in and out of a randomized trial and illustrates it by using the PBIS data. In particular, Section 2 describes previous work in methods to assess or enable generalizability. Section 3 then proposes two diagnostic measures that use propensity scores to help to quantify how similar the subjects in a trial are to the target population. Section 4 applies those measures to the motivating example of the PBIS programme, and Section 5 concludes.

The programs that were used to analyse the data can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Previous work assessing generalizability

To this point, the emphasis of research has generally been on internal validity—obtaining unbiased effect estimates for the participants in a trial. Less attention has been paid to external validity—addressing whether those participants are representative of the target population and whether the effects are generalizable (Imai *et al.*, 2008). The following two sections describe the methods that have been used to assess generalizability, first in terms of study design strategies and then in terms of data analysis techniques.

2.1. Study design

To provide a framework for thinking about these issues, Imai *et al.* (2008) decomposed the estimation error in the estimate of a population treatment effect (Δ) into components due to sample selection and to treatment assignment:

$$\Delta = (\Delta_{\mathrm{S}X} + \Delta_{\mathrm{S}U}) + (\Delta_{\mathrm{T}X} + \Delta_{\mathrm{T}U}),$$

where S refers to bias due to sample selection and T refers to bias due to treatment selection. The subscript X refers to observed variables and U to unobserved. Different study designs focus on different quantities. For example, randomized experiments have $\Delta_{\mathrm{T}X} = \Delta_{\mathrm{T}U} = 0$ but may have larger $\Delta_{\mathrm{S}X}$ and $\Delta_{\mathrm{S}U}$ than do observational studies. Observational study methods such as propensity score matching (Stuart, 2010) focus on reducing $\Delta_{\mathrm{T}X}$, and sensitivity methods such as in Rosenbaum (2002) assess the potential effect of $\Delta_{\mathrm{T}U}$ on study conclusions.

Relatively less attention has been paid to the size of $\Delta_{\mathrm{S}X}$ and $\Delta_{\mathrm{S}U}$ in randomized experiments. Standard methods that make qualitative arguments regarding the generalizability of results

assume that the results from the trial directly carry over to the population: that $\Delta_{S_X} = \Delta_{S_U} = 0$. In this paper we focus on methods to assess the amount of sample selection bias due to observed covariates, Δ_{S_X}. Although there may still be bias due to Δ_{S_U}, the methods that are discussed here at least provide a way to reduce bias in Δ due to Δ_{S_X}.

One of the most straightforward ways of ensuring the generalizability of results from randomized trials is to enrol in the trial a representative sample from the target population. However, only a handful of studies have used random assignment of a fully representative (e.g. random) sample from a population to estimate programme effects (Cook, 2007). Examples include the national evaluations of 'Upward bound' (US Department of Education, 2009) and 'Head start' (US Department of Health and Human Services, 2010), although even that 'Head start' evaluation excluded certain centres, such as those that were underenrolled and those serving Native American populations. Increasing attention has been given to practical clinical trials, which aim to enrol a very large and diverse sample of patients, from a range of settings (Peto *et al.*, 1995; Insel, 2006). However, those trials require large amounts of time and money and are not always feasible. There has also been some discussion of purposively sampling units that are either heterogeneous (to reflect the range of units that are in the target population) or that are typical of that population (Shadish *et al.*, 2002), but those ideas seem to have been rarely used in practice, at least in a formal way.

2.2. Study analysis

Another strategy is to use existing data to assess the generalizability of existing studies, which is the approach that we take here. Other methods in this area include meta-analysis (Hedges and Olkin, 1985; Sutton and Higgins, 2008), cross-design synthesis (Prevost *et al.*, 2000) and the confidence profile method (Eddy *et al.*, 1992). Many of these approaches aim to model treatment effects as a function of study parameters, such as randomized *versus* non-randomized, and the explicit inclusion–exclusion criteria, and they generally rely on having a relatively large set of studies to include in the analysis. Unfortunately there is often only one or two studies from which conclusions can be drawn. In addition, little attention is usually paid to the types of participants who are enrolled in the various studies included, and how variation in their characteristics may affect the results.

Perhaps the most common way of generalizing results to a target population is through post-stratification, which reweights the effects based on population distributions. As a simple example, imagine a target population with 50% males and 50% females, but a randomized trial that had 20% males and 80% females. A simple post-stratification would estimate effects separately by gender and then average the male and female effects by using the population proportions (50–50). Post-stratification can be very effective when there are only a small number of variables to control for, but it is infeasible when there are many (or continuous) variables, leading to a very large number of post-stratification cells. Frangakis (2009) discussed a more complex scenario for post-stratification, where generalizability also depends on post-treatment variables. Post-stratification is closely related to methods that model treatment-by-covariate interactions to investigate whether effects vary across individuals (e.g. Rothwell (2005b) and Wang *et al.* (2007)). In fact investigation of subgroup effects and effect heterogeneity is a crucial step in determining what covariates are crucial to control for in the methods that we propose, as discussed further below.

Weisberg *et al.* (2009) posited a simple model to account for differences between a trial sample and a population due to inclusion or exclusion criteria, for a setting with a binary outcome. They provided formulae for the amount of bias that may be created and showed that, depending on whether high risk patients are particularly included or excluded, the estimated effect may

change considerably, or even reverse sign. In work that is probably most similar to that presented here, Greenhouse *et al.* (2008) provided a case-study of assessing generalizability, comparing the characteristics of paediatric participants in randomized trials of antidepressants to the general population of children and adolescents. That work represented an important advance in raising this issue, in the context of an important policy question regarding antidepressants and suicidal tendencies.

3. Using propensity scores to assess generalizability

3.1. Formal setting

We consider a setting where a randomized trial has been conducted to estimate the effect of a programme P relative to a control condition C on a sample of participants Ψ of size n. By 'programme' we mean any intervention of interest, whether preventive or a treatment for a particular disorder or disease. The participants in Ψ may be individuals or they may be at a higher level, such as communities or schools, as in the case of the PBIS programme. In the trial the programme P has been randomly assigned to participants in Ψ, forming a programme 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 programme P in a target population of size N, which is represented by Ω, where Ψ is a subset of Ω. We refer to Ψ as the 'sample' and Ω as the 'population'. In the PBIS example, Ψ consists of the 37 schools in the effectiveness trial; Ω consists of all public elementary schools in the state. We assume that for all participants in Ω (or a representative sample of them) we observe a set of background characteristics X, which describe both the participants themselves and their broader contexts. In the PBIS study, X consists of characteristics such as test scores, enrolment and demographics.

For subject i we denote membership in the randomized trial sample by S_i, T_i indicates membership in the treatment *versus* control group (which is only defined for those with $S_i = 1$) and $Y_i(1)$ and $Y_i(0)$ are the potential outcomes under treatment and control respectively (Rubin, 1977). Following the notation in Imai *et al.* (2008), the treatment effect for individual i is the difference in potential outcomes, $Y_i(1) - Y_i(0)$, although results could be extended to other functions of the potential outcomes, such as their ratio. The standard intent-to-treat estimates from a randomized trial, such as a difference in means of the outcome in the treated and control groups, yields an unbiased estimate of the sample average treatment effect SATE:

$$\text{SATE} = \frac{1}{n} \sum_{i \in \{S_i = 1\}} Y_i(1) - Y_i(0).$$

However, our estimand of interest is the population average treatment effect PATE:

$$\text{PATE} = \frac{1}{N} \sum_{i=1}^{N} Y_i(1) - Y_i(0).$$

When the treatment effect is constant PATE = SATE, but that will generally not be so. Although PATE is a clearly defined measure of impact in a given population, in many research settings the target population changes over time or space. In such cases, there may be more than one PATE that is of interest. For a simple setting with one effect modifier, Cole and Stuart (2010) have given an equation for the bias, which depends on

(a) the proportion of the population that is not sampled,
(b) the heterogeneity in the treatment effects,

(c) the prevalence of the effect modifier in the population and
(d) the strength of the association between the effect modifier and sample selection.

3.2. Key assumptions

We make three primary assumptions. Assumption 1 is that, given X, all subjects in the population have some probability of being selected for the trial:

$$0 < P(S_i = 1|X_i) < 1 \qquad \text{for all } X_i.$$

Assumption 2 is that there are no unmeasured variables that are related to both sample selection and the treatment effect ($E(\Delta_{S_U}) = 0$), which we term 'unconfounded sample selection' (see also Cole and Stuart (2010)). This assumption is similar to the assumption of ignorable treatment assignment in observational studies (Rosenbaum and Rubin, 1983a) or the 'missingness at random' assumption with respect to missing data (Rubin, 1976). Formally, assumption 2 says that sample selection is independent of the potential outcomes, given the observed covariates:

$$S \perp [Y(0), Y(1)]|X.$$

Assumption 3 is that treatment assignment is random (and hence independent of the potential outcomes) and independent of sample selection, given the observed covariates:

$$T \perp [S, Y(0), Y(1)]|X.$$

The combination of assumptions 1 and 3 also implies that each subject has a positive probability of receiving the treatment, which is comparable with the assumption of 'strongly ignorable treatment assignment' in Rosenbaum and Rubin (1983a).

The validity of assumption 1 depends on the definition of the target population. In settings where some individuals in the initial target population would never receive the treatment of interest (e.g. a targeted intervention that is only given to at-risk students who have already exhibited some problem behaviours), the target population should be redefined to include only those individuals to whom the treatment may be given. Assumption 3 is met in randomized trials where the random assignment is done after the sample selection, and where treatment assignment depends only on observed characteristics X. It thus arguably is not an assumption *per se*; we include it here for completeness. Assumption 2 is arguably the most difficult to meet and relies on having all potential moderators of the treatment effect measured. This assumption is discussed further below.

Given these three assumptions we discuss two diagnostics for generalizability, both based on the propensity score: first, the average propensity score difference between the sample and the population, and, second, the use of propensity score methods to compare observed and predicted outcomes under control for the population.

3.3. Propensity score distance as a measure of similarity

The first diagnostic tool that we propose is the propensity score distance between the participants in the trial and the target population, as a way to summarize their similarity. Here, the propensity scores model the probability of being in the randomized trial. The propensity score is typically defined as the probability of receiving some programme (or 'treatment') *versus* a comparison condition, given a set of observed baseline characteristics (Rosenbaum and Rubin, 1983a). Propensity score matching, subclassification or weighting can help to ensure that the programme and comparison subjects being compared 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. In this way, propensity scores attempt to replicate a randomized experiment in the sense of comparing subjects who did and did not receive the treatment who have no systematic differences on the observed covariates (Ho *et al.*, 2007; Stuart, 2010).

Here, to summarize differences between the trial sample and the target population, the propensity score will model membership in the randomized trial sample, rather than receipt of the treatment. In particular, we use a logistic regression model of the probability of being in the randomized trial (S) with the covariates X as predictors:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki},$$

where $p_i = P(S_i = 1|X_i)$. We use $\hat{p}_i$ to denote the estimated probability of sample selection for subject i. In a connection to the discriminant function, these propensity scores serve as the scalar summary of the covariates that distinguishes the most between the trial participants and the target population (Rubin and Thomas, 1992). They thus can provide a summary measure of the similarity (or dissimilarity) of the trial participants and the population: just as propensity scores can be used to identify when treatment and control groups are too far apart for reliable causal inference (Rubin, 2001), they can be used also to identify when a sample is too different from a population of interest to yield reliable generalizations.

We define the propensity score difference between the trial sample and population (Δ_{p}) as the difference in average propensity scores between those who are in the trial and those who are not in the trial:

$$\Delta_{\mathrm{p}} = \frac{1}{n} \sum_{i \in \{S_i = 1\}} \hat{p}_i - \frac{1}{N-n} \sum_{i \in \{S_i = 0\}} \hat{p}_i.$$

If the sample is actually a (very large) random sample from the population, then we would expect essentially no difference in the mean propensity scores between those sampled and not sampled. In that case, $E(p_i|S_i = 1) = E(p_i|S_i = 0)$ and $E(\Delta_{\mathrm{p}}) = 0$. As we shall see later in the PBIS example, in finite samples, there is likely to be a small positive value for Δ_{p}, reflecting small chance differences between the trial participants and the population. When there are systematic differences between the trial sample and population we shall expect that these means will be quite different. In the standard propensity score context of observational studies, simulation studies and theoretical approximations that were originally developed in the context of matching within propensity score calipers (Cochran and Rubin, 1973; Rubin, 1973) have indicated that propensity score means that differ by more than 0.25 standard deviations indicate a large amount of extrapolation and heavy reliance on the models being used for estimation (Ho *et al.*, 2007; Stuart, 2010), although this is by no means a set criterion, and in fact some researchers recommend a more stringent criterion of 0.1 (Mamdani *et al.*, 2005).

3.4. Using propensity score methods to match the control group to the population and to compare predicted and observed outcomes

A second way in which we can use the propensity scores to assess generalizability is by using propensity score methods to make the control group look like the target population and to compare the predicted outcomes under control with the outcomes that are actually observed in the population. We discuss three methods here: IPTW, full matching and subclassification. All three of these methods can be thought of as weighting the control group to the population; the

methods vary in the coarseness of the weights, with subclassification the coarsest and IPTW the finest. In fact, IPTW can be thought of as an extreme where the number of individuals and subclasses goes to ∞ (Rubin, 2001).

IPTW methods give each individual their own weight, which is calculated as the inverse propensity scores, i.e., in our setting, the inverse probability of being in the sample: $w_i(X_i) = 1/\hat{p}_i(X_i)$. These are conceptually similar to the weights that are used in non-response adjustments in survey sampling (Kalton and Flores-Cervantes, 2003). This weighting forms a pseudo-population with characteristics that are similar to those of the target population. If no one in the population was receiving the intervention of interest, then, if the weights are effective, the weighted control group outcomes should be similar to the outcomes that are observed in the target population. Mathematically, we can extend results in Horvitz and Thompson (1952) and Lunceford and Davidian (2004), where the expectations are taken over the target population:

$$
\begin{aligned}
E\left[\frac{S(1-T)Y}{w(X)\{1-e(X)\}}\right] &= E\left(E\left[\frac{\mathbf{1}(S=1)\{1-\mathbf{1}(T=1)\}Y}{w(X)\{1-e(X)\}}|Y,X\right]\right) \\
&= E\left(E\left[\frac{\mathbf{1}(S=1)\{1-\mathbf{1}(T=1)\}Y(0)}{w(X)\{1-e(X)\}}|Y(0),X\right]\right) \\
&= E\left[\frac{Y(0)}{w(X)\{1-e(X)\}}E\{\mathbf{1}(S=1)\}\{1-\mathbf{1}(T=1)|Y(0),X\}\right] \\
&= E\left[\frac{Y(0)}{w(X)\{1-e(X)\}}P(S=1|X=x)\{1-P(T=1|X=x)\}\right] \\
&= E\{Y(0)\}.
\end{aligned}
$$

In this expression $e(X)$ reflects the probability of treatment assignment, $e(X) = P(T=1|X)$, which is known in a randomized experiment. The above equations show that the probability of treatment assignment and trial participation-weighted control group mean will be an unbiased estimate of the population potential outcome under control, given the assumptions that were detailed above. In this way we can use the similarity of the weighted control group means to the population means as a diagnostic for how well the generalization is likely to work.

However, a concern about IPTW methods is that the results can be somewhat unstable, especially if there are extreme weights, and the method is more sensitive to the specification of the propensity score model than are other propensity score approaches (Kang and Schafer, 2007). We thus also consider two other methods of reweighting the control group to resemble the population, which use coarser weights. At the other extreme, subclassification methods form a relatively small (e.g. 5–10) number of subclasses and group individuals with similar propensity score values (e.g. by the quintiles of the propensity score distribution). However, subclassification approaches can suffer from having too few subclasses and thus insufficient bias reduction (Lunceford and Davidian, 2004; Stuart, 2010).

We thus also consider a third approach: full matching (Hansen, 2004; Stuart and Green, 2008), which can be thought of as a compromise between IPTW and subclassification (Stuart, 2010). Full matching forms a relatively large number of subclasses, where in our use each subclass will have at least one member of the sample and at least one member of the target population, but the ratio of sample to population members in each subclass can vary. The subclasses reflect the fact that some areas of the propensity score space will have relatively few sample members and many population members, whereas other areas will have relatively few population members and many sample members. Full matching has been shown to be optimal in terms of reducing propensity score differences within subclasses (Rosenbaum, 1991).

For both subclassification and full matching the control group members in the trial are given weights that are proportional to the number of population members in their subclass. For example, sample members in a subclass with two sample members and 10 population members would receive weights proportional to 5 (10/2), whereas sample members in a subclass with 10 sample members and two population members would receive weights that are proportional to 0.2 (2/10). For details on the construction of the weights following full matching see Stuart and Green (2008).

4. Applying methods to 'Positive behavioral interventions and supports' study

We now apply the diagnostic tools that were described above to the group-randomized trial of the PBIS programme, which is a schoolwide prevention programme that aims to improve school climate by creating improved systems and procedures that promote positive change in staff and student behaviours (Sugai and Horner, 2006). It is being widely disseminated by the US Department of Education and many state governments. By 2010, over 10000 schools across the USA, representing approximately 10% of all US public schools, were implementing the PBIS programme (see `http://www.pbis.org`) (Technical Assistance Center on Positive Behavioral Interventions and Supports, 2010). Because the intervention operates at the school level, the unit of analysis is the school.

We combine information from two data sets to illustrate the use of propensity scores for assessing generalizability. First, the randomized effectiveness trial of the universal system of schoolwide PBIS, which was called 'Project target' (PT), began in 2002 among a sample of 37 Maryland public elementary schools that volunteered for the study (Bradshaw *et al.*, 2009). Those 37 schools were randomized to treatment and control in two years (2002, 2003); for our illustrative purposes here we pool the two years. Second, we have longitudinal data on all public elementary schools across Maryland, from 1993 through to 2007. This provides the population data that are necessary to estimate the probabilities of participation in the trial and to compare the schools in the trial with schools across the state.

State level co-ordinated training is required to implement the PBIS programme. The state of Maryland has a mechanism for training schools in the PBIS programme, which was used to train both the PBIS schools in the trial and non-trial schools that chose to implement the PBIS programme (Barrett *et al.*, 2008). Because some elementary schools were implementing the programme outside the trial, as the target population we consider the 717 elementary schools in the state that had not implemented it by 2006 and that were not participating in the PT trial; subsequent use of the term 'state population' refers to this subsample of the full state population. Excluding the schools participating in the trial from the state population of interest increases clarity and precision.

We consider variables that were measured in 2002 as pretrial covariates and examine outcomes measured in 2003 and later. Observed characteristics of the schools (both in the trial and state wide) include characteristics of the students (e.g. the percentage of students classified as special education, the percentage who qualify for free or reduced price meals and average mathematics and reading test scores), as well as of the schools themselves (e.g. enrolment). Schools in the PBIS trial are somewhat different from the population of elementary schools across Maryland on these observed characteristics (Table 1). In 2002, the schools that were enrolled in the trial had somewhat lower test scores, more students eligible for title 1 (a measure of poverty) and higher rates of suspension than other schools across the state.

To summarize these differences, a propensity score (logistic regression) model was fitted predicting membership in the PT trial given the set of characteristics in Table 1. Fig. 1 shows the

Table 1. Baseline characteristics of schools in the PT trial and schools across the state of Maryland†

Characteristic (2002)	*Results for PT schools*		*Results for non-PT schools*		*p-value of difference*
	Mean	*Standard deviation*	*Mean*	*Standard deviation*	
Total enrolment	485	150	480	177	0.85
Attendance rate (%)	95.3	0.7	95.4	1.5	0.80
% students Caucasian	60.3	31.7	53.8	34.0	0.23
% students eligible for free or reduced price meals	39.7	20.0	36.2	27.5	0.31
% students eligible for title 1	47.3	49.6	26.3	41.4	0.02
% students in special education	13.8	5.6	15.2	15.4	0.21
3rd-grade mathematics test	27.4	15.2	32.1	20.5	0.08
3rd-grade reading test	32.9	16.4	34.5	20.4	0.57
5th-grade mathematics test	44.6	18.6	51.3	29.9	0.05
5th-grade reading test	54.2	17.9	53.2	26.0	0.75
Trend in 3rd-grade mathematics scores	−19.2	18.9	−15.9	16.6	0.31
Trend in 3rd-grade reading scores	−12.3	18.6	−11.9	14.6	0.90
Trend in 5th-grade mathematics scores	−13.4	20.9	−11.9	18.4	0.66
Trend in 5th-grade reading scores	1.3	19.1	−1.9	16.8	0.33
% of students suspended	6.3	4.5	4.5	5.1	0.03
Sample size	37		680		

†All variables were measured in 2002. Test score variables reflect the percentage of students scoring in the 'advanced' or 'proficient' ranges on the Maryland state standardized test. The trend shows a change from 2000 to 2002. *p*-values are shown from *t*-tests or χ^2-tests, as appropriate.

distribution of propensity scores among schools across the state and for the schools in the PT trial. In general there is overlap of the propensity scores, with many of the trial schools in the range of propensity scores with high density among the schools across the state, but also with a number of the trial schools with relatively large propensity scores. We can also quantify this, in that the difference in average propensity scores between the schools in the trial and those across the state (Δ_p) is 0.055. The standardized difference (standardized by the standard deviation of the propensity score) is 0.73, which is a substantial difference, and a size indicated by Rubin (2001) to lead to unreliability of standard regression modelling because of the resulting extrapolation.

We can also compare these differences with what would be expected in repeated random draws of the same size. This allows us to determine what size propensity score difference we would expect if the schools in the trial were in fact selected randomly from the population. Fig. 2 shows the distribution of the difference in mean propensity scores between sampled and unsampled schools, given repeated samples of size 37 drawn from the population of Maryland public elementary schools that had not implemented the PBIS programme by 2006. Because the propensity score is defined relative to a particular sample, the propensity scores themselves are recalculated within each random sample. Whereas a propensity score distance of approximately 0.02 would be expected, our observed difference of 0.055 would happen in only 24 of 1000 samples randomly drawn from the population. Similarly, whereas a standardized difference of size 0.5 would be expected, only three of the 1000 samples that we drew had a standardized difference as large as our observed value of 0.73.

We then used the three propensity score adjustment methods that were described above (IPTW, full matching and subclassification) to match the control group in the PT trial to

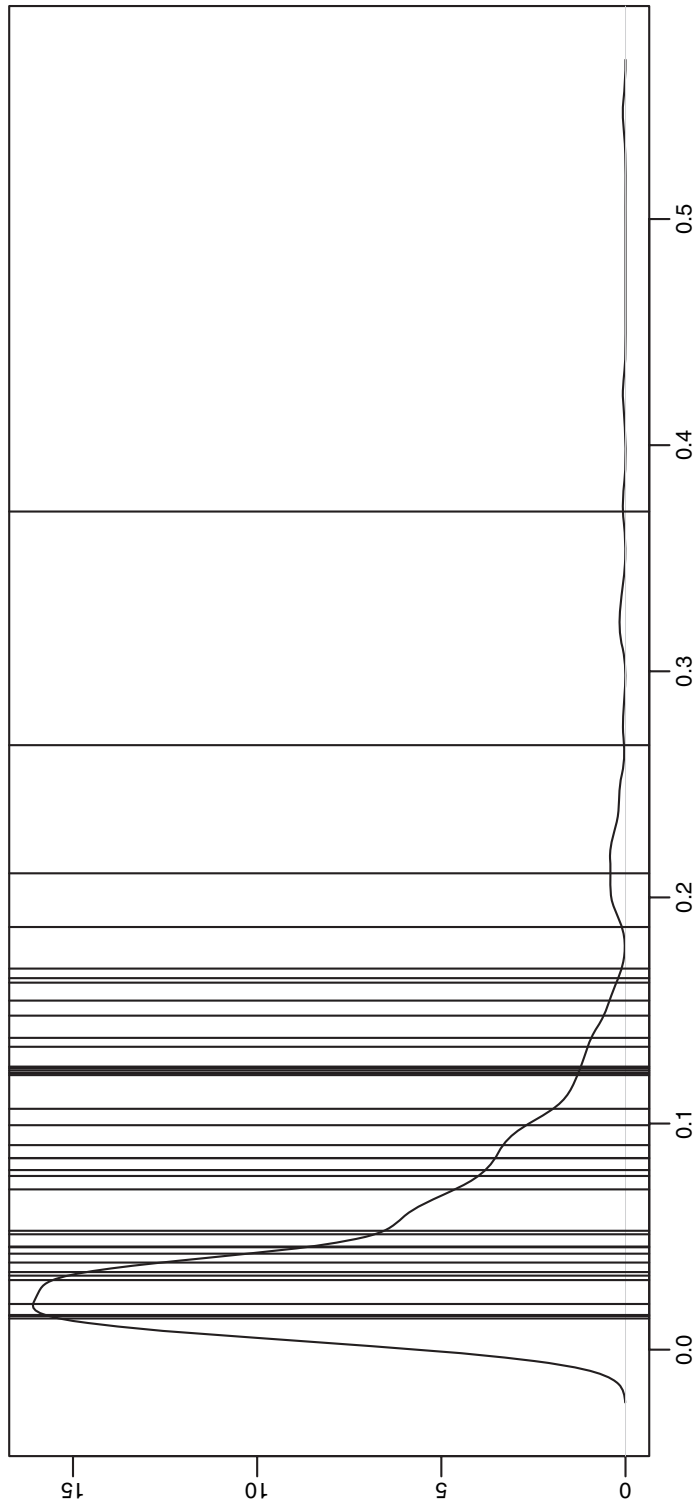

Fig. 1. Distribution of propensity scores among the schools across the state (——) and schools in the PT trial (|): the state population consists of all elementary schools across the state of Maryland not implementing the PBIS programme and not enrolled in the trial

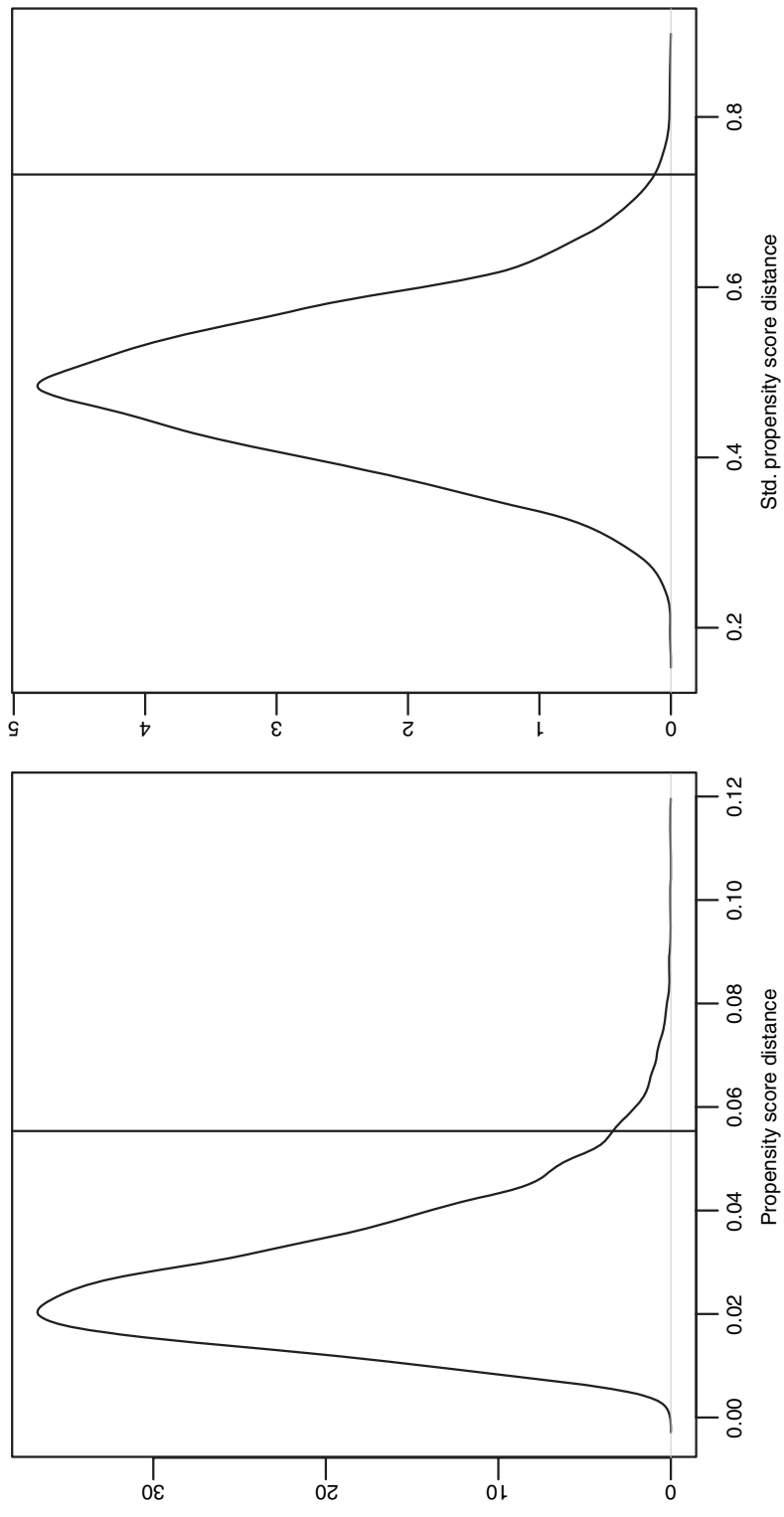

Fig. 2. Distribution of propensity score distances between sampled and unsampled schools, where samples of size 37 were repeatedly drawn from the population of elementary schools in Maryland (|, value observed for the PT trial schools): (a) simple differences (sampled minus unsampled); (b) standardized difference, standardized by the standard deviation of the propensity scores

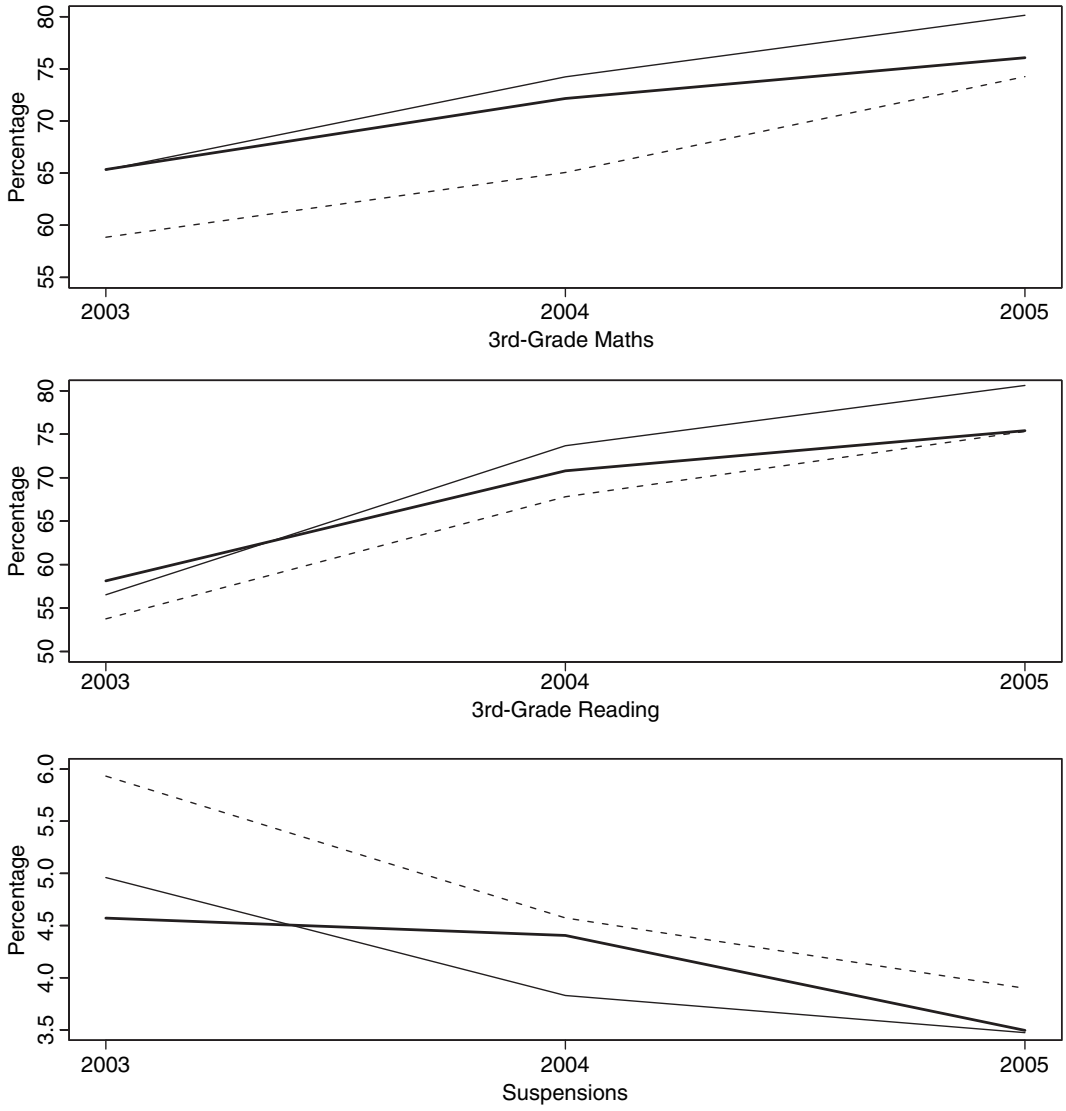

Fig. 3. Observed and predicted outcome values for schools across the state of Maryland (———, observed state averages, where the state population refers to schools across the state not implementing the PBIS programme and not enrolled in the trial; ------, average for control schools in the PT trial; ———, weighted average for control schools in the trial, with weights calculated from full matching): for mathematics and reading scores, the numbers shown are percentages of children scoring 'proficient' or 'advanced' on the standardized test; numbers shown for suspensions are the percentages of students suspended in a school year; for all three outcomes, the weighted average tracks the state mean much more closely than the observed average among control schools in the trial

the state population. The results were broadly similar for the three approaches and thus we show only the results for full matching, which can be considered the intermediate approach. Fig. 3 thus illustrates the second diagnostic tool, which is to examine the comparability of the weighted control group means with the values that were observed in the population. Despite the differences seen between the trial and non-trial schools above, it appears that the control schools in the trial reflect what was happening across the state as a whole, when weighted up to

represent the population. We used three outcomes for this illustration: the percentage of third-grade students scoring proficient or higher on the yearly statewide reading and mathematics examinations and the percentage of students who were suspended during the school year. The bold curve shows the average value for each outcome, averaged across all schools in the state that were not implementing the PBIS programme and not in the trial, over the time period from 2003 to 2005. The broken curve shows the same average, but calculated by using the control schools in the trial. The large distance between the broken curve and the bold curve illustrates the overall difference between the schools in the trial and those in the state: on average, the schools in the trial had lower test scores and higher rates of suspension than those across the state. The thin curve shows the average for the control schools in the trial, but weighted by using weights calculated from propensity score full matching. For all three outcomes, and especially the test scores, the trial control schools' weighted average tracks the true state mean quite closely, which is seen by the similarity of the two full curves in each panel of Fig. 3. This is particularly true for outcomes in 2003–2004, which is expected given that the propensity scores were estimated by using variables that had been measured in 2002 and before. Interestingly, using IPTW seemed to track the population outcomes slightly better than full matching, whereas subclassification tracked the population outcomes slightly less well. Future work should further compare these approaches and their use in generalizing trial results. These results indicate that, despite the apparently large differences on the pretreatment covariates, when weighted appropriately, the schools in the trial may help us to learn about population effects across the state of the PBIS programme on third-grade mathematics and reading tests and rates of suspension.

5. Discussion

Propensity scores offer a promising way to assess the similarity between a trial sample and a target population of interest. However, there is still much work remaining in this area. The methods that were presented here assume that individual level data are available for the population, or at least for a representative sample from that population. This type of data is becoming more readily available, through nationally representative data sets or through administrative data sets such as Medicare claims, Veteran's Administration records files or health system administrative data sources. Similarly, the No Child Left Behind Act requires states to collect and keep school records data on academic performance and discipline for reporting purposes. However, in cases where individual (or school) level population data may not be available, Cole and Stuart (2010) have presented an alternative approach that uses only summary statistics on the population of interest.

An important question for this work regards variable selection and model estimation. These issues may be informed by similar work in the broader propensity score literature (e.g. Brookhart *et al.* (2006)); however, it will be important to consider whether the implications are different in this setting. For example, it will probably be especially important to include strong predictors of the outcome, and particularly moderators of the treatment effects, in the propensity score model. Because of this, it is also important for current trials to investigate effect heterogeneity and which covariates are effect modifiers. A related issue is how to weight the characteristics that are observed. By default, propensity scores effectively weight each characteristic by how predictive it is of membership in the trial sample. However, researchers are likely to have prior information on which characteristics moderate the treatment effects and thus are particularly important to control for. Future work will develop methods that prioritize these key characteristics by giving them more emphasis in the summary measure. One possible approach is to combine propensity scores with another multivariate distance (such as the Mahalanobis distance)

calculated by using those key characteristics (Rubin and Thomas, 2000). A second possible approach is to combine propensity scores with the prognosis score methods that have recently been developed by Hansen (2008). Prognosis scores effectively weight each characteristic by how predictive it is of the outcome under control, $Y(0)$.

In many settings it is likely that both individual and contextual level factors moderate the effects of interventions. For example, the effects of school-based smoking prevention programmes may be moderated by both individual characteristics such as gender and race, but also by the characteristics of the school and community. In the current study we have dealt with this by incorporating both individual level measures as well as characteristics of the schools. In other cases, where individual level data are available, a more direct way of doing so would be to use a multilevel or hierarchical framework, in which the relationship between participation in the trial and the individual and context level characteristics are modelled at separate levels: one for individuals and one for the context. This is discussed in the context of randomized experiments in Brown *et al.* (2008) and Hong and Raudenbush (2006).

The methods that were described in this paper assume that all the effect moderators are observed (unconfounded sample selection). This is a crucial assumption, and it is important to assess the validity of it. In the case of the PBIS programme, there has been very little research into the effectiveness of the intervention (Bradshaw *et al.*, 2010), and in fact the Maryland trial was only the second randomized trial of the PBIS programme in the country (the other is described in Horner *et al.* (2009)). Therefore, relatively little is known about potential school level moderators of the effects of the PBIS programme. The theory of this intervention, and of school-based interventions in general, leads us to believe that we probably observe most of the major variables that may affect the effectiveness of the PBIS programme (e.g. academic achievement and size of school). However, some important variables, such as the schools' organizational capacity to implement the programme (Bradshaw *et al.*, 2009), the Principals' support for it (Kam *et al.*, 2003) or the reasons why the Principals volunteered the school to participate in the trial, are missing from these analyses. Although some of these data are available for the schools in the trial, unfortunately none are available state wide. A potential consequence of these possible unobserved confounders is that the methods that are described in this paper may be more appropriate for some outcomes than for others. For example, we repeated the analyses for fifth-grade test scores and found that there were still substantial differences between the weighted control schools' outcomes and the average outcomes across the state. One potential explanation for this varied performance across grades is possible unobserved differences in problems of discipline between the schools in the trial and those across the state. Schools that volunteered to participate in the trial may have done so in part because of relatively high disciplinary problems, which tend to manifest themselves more in the later elementary school years (Koth *et al.*, 2009); the Principals of the participating schools may have felt more need to participate in the trial to have the possibility of receiving the PBIS programme. Thus, even when weighted by using the characteristics in Table 1, the fifth-graders in the trial schools may have been more different from those across the state, especially compared with third-graders. This would be an example of an unmeasured characteristic that differs between the sample and population, which in particular may impact fifth-grade scores more than third-grade scores. This may in part also be why the results for suspensions look somewhat worse than for third-grade test scores (Fig. 3). The diagnostics that are presented here can help to determine when the weights are sufficient for generalization, or when unobserved variables are likely to cause a problem, as seen for the fifth-grade test scores. Future work will also investigate methods to assess the sensitivity of effect estimates to an unobserved moderator, along the lines of Rosenbaum and Rubin (1983b).

As increasingly more high quality effectiveness trials are carried out, the clear next research questions will involve external validity and generalizing the results from those trials. Some recent work has started to investigate weighting methods to generalize results to target populations, but diagnostics are first needed to help to determine when such generalization is reasonable. The methods that were proposed here provide a first step towards assessing the similarity of participants in a trial to those in a population, allowing researchers to begin to examine the extent to which the results that are seen in trials may generalize more broadly. Assuming that these diagnostics indicate that it is safe to proceed, future work should expand these approaches to generate effect estimates for target populations of interest.

Acknowledgements

This research was supported by the National Institute of Mental Health (K25 MH083846, Principal Investigator Stuart; 1 R01 MH67948-1A1, Principal Investigator Leaf), the Centers for Disease Control (R49/CCR318627, Principal Investigator Leaf) and the Institute of Education Sciences (R305A090307, Principal Investigator Bradshaw).

References

Barrett, S., Bradshaw, C. and Lewis-Palmer, T. (2008) Maryland state-wide PBIS initiative: systems, evaluation, and next steps. *J. Pos. Behav. Intervens*, **10**, 105–114.

Bradshaw, C. P., Koth, C. W., Thornton, L. A. and Leaf, P. J. (2009) Altering school climate through school-wise Positive Behavioral Interventions and Supports: findings from a group-randomized effectiveness trial. *Prevn Sci.*, **10**, 100–115.

Bradshaw, C., Mitchell, M. and Leaf, P. (2010) Examining the effects of Schoolwide Positive Behavioral Interventions and Supports on student outcomes: results from a randomized controlled effectiveness trial in elementary schools. *J. Pos. Behav. Intervens*, **12**, 133–148.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J. and Sturmer, T. (2006) Variable selection for propensity score models. *Am. J. Epidem.*, **163**, 1149–1156.

Brown, C. H., Wang, W., Kellam, S. G., Muthen, B., Petras, H., Toyinbo, P., Poduska, J., Ialongo, N., Wyman, P. A., Chamberlain, P., Sloboda, Z., MacKinnon, D. P., Windham, A. and the Prevention Science Methodology Group (2008) Methods for testing theory and evaluating impact in randomized field trials: intent-to-treat analyses for integrating the perspectives of person, place, and time *Drug Alc. Depend.*, **95**, S74–S104.

Campbell, D. T. (1957) Factors relevant to the validity of experiments in social settings. *Psychol. Bull.*, **54**, 297–312.

Cochran, W. G. and Rubin, D. B. (1973) Controlling bias in observational studies: a review. *Sankhya* A, **35**, 417–446.

Cole, S. R. and Stuart, E. A. (2010) Generalizing evidence from randomized clinical trials to target populations: the ACTG-320 trial. *Am. J. Epidem.*, **172**, 107–115.

Cook, T. D. (2007) Evidence-based practice: where do we stand? In *Proc. 20th A. Res. Conf. System of Care for Children's Mental Health: Expanding the Research Base*. Tampa: University of South Florida. (Available from `http://rtckids.fmhi.usf.edu/rtcconference/20thconference/iding.cfm`.)

Couper, M. P. and Miller, P. V. (2008) Web survey methods: introduction. *Publ. Opin. Q.*, **72**, 831–835.

Eddy, D., Hasselblad, V. and Shachter, R. (1992) *Meta-analysis by the Confidence Profile Method: the Statistical Synthesis of Evidence*. New York: Academic Press.

Flay, B. R. (1986) Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Prevn Med.*, **15**, 451–474.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., Moscicki, E. K., Schinke, S. and Valentine, J. (2005) Standards of evidence: criteria for efficacy, effectiveness and dissemination. *Prevn Sci.*, **6**, 151–175.

Frangakis, C. E. (2009) The calibration of treatment effects from clinical trials to target populations. *Clin. Trials*, **6**, 136–140.

Glasgow, R. E., Nelson, C. C., Strycker, L. A. and King, D. E. (2006) Using RE-AIM metrics to evaulate diabetes self-management support interventions. *Am. J. Prev. Med.*, **30**, 67–73.

Greenhouse, J. B., Kaizar, E. E., Kelleher, K., Seltman, H. and Gardner, W. (2008) Generalizing from clinical trial data: a case study of the risk of suicidality among pediatric antidepressant users. *Statist. Med.*, **27**, 1801–1813.

Grodstein, F., Clarkson, T. and Manson, J. (2003) Understanding the divergent data on post-menopausal hormone therapy. *New Engl. J. Med.*, **348**, 645–650.

Haneuse, S., Schildcrout, J., Crane, P., Sonnen, J., Breitner, J. and Larson, E. (2009) Adjustment for selection bias in observational studies with application to the analysis of autopsy data. *Neuroepidemiology*, **32**, 229–239.
Hansen, B. B. (2004) Full matching in an observational study of coaching for the SAT. *J. Am. Statist. Ass.*, **99**, 609–618.
Hansen, B. B. (2008) The prognostic analogue of the propensity score. *Biometrika*, **95**, 481–488.
Hedges, L. V. and Olkin, I. (1985) *Statistical Methods for Meta-analysis*. Burlington: Academic Press.
Hernán, M., Alonso, A., Logan, R., Grodstein, F., Michels, K., Willett, W., Manson, J. and Robins, J. (2008) Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease (with discussion). *Epidemiology*, **19**, 766–779.
Ho, D. E., Imai, K., King, G. and Stuart, E. A. (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.*, **15**, 199–236.
Hong, G. and Raudenbush, S. W. (2006) Evaluating kindergarten retention policy: a case study of causal inference for multilevel observational data. *J. Am. Statist. Ass.*, **101**, 901–910.
Horner, R., Sugai, G., Smolkowski, K., Eber, L., Nakasato, J., Todd, A. and Esperanza, J. (2009) A randomized, wait-list controlled effectiveness trial assessing school-wide Positive Behavior Support in elementary schools. *J. Pos. Behav. Intervens*, **11**, 133–144.
Horvitz, D. and Thompson, D. (1952) A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Ass.*, **47**, 663–685.
Imai, K., King, G. and Stuart, E. A. (2008) Misunderstandings between experimentalists and observationalists about causal inference. *J. R. Statist. Soc.* A, **171**, 481–502.
Insel, T. R. (2006) Beyond efficacy: the STAR∗D trial. *Am. J. Psychiatr.*, **163**, 5–7.
Institute of Medicine (2006) *Improving the Quality of Health Care for Mental and Substance-use Conditions*. Washington, DC: National Academies Press.
Kalton, G. and Flores-Cervantes, I. (2003) Weighting methods. *J. Off. Statist.*, **19**, 81–97.
Kam, C., Greenberg, M. and Walls, C. (2003) Examining the role of implementation quality in school-based prevention using the PATHS curriculum. *Prevn Sci.*, **1**, 55–63.
Kang, J. D. and Schafer, J. L. (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.*, **22**, 523–539.
Koth, C., Bradshaw, C. and Leaf, P. (2009) Teacher Observation of Classroom Adaptation-Checklist (TOCA-C): development and factor structure. *Measmnt Evaln Counsel. Devlpmnt*, **42**, 15–30.
Lunceford, J. K. and Davidian, M. (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statist. Med.*, **23**, 2937–2960.
Mamdani, M. M., Sykora, K., Li, P., Normand, S.-L. T., Streiner, D. L., Austin, P. C., Rochon, P. A. and Anderson, G. M. (2005) Reader's guide to critical appraisal of cohort studies: 2, assessing potential for confounding. *Br. Med. J.*, **330**, 960–962.
National Institute of Mental Health (1999) Bridging science and service: a report by the NIMH Council's clinical treatment and services research workgroup. *Technical Report*. National Institute of Mental Health, Bethesda. (Available from `http://www.nimh.nih.gov/publicat/nimhbridge.pdf`.)
Pan, Q. and Schaubel, D. E. (2009) Evaluating bias correction in weighted proportional hazards regression. *Liftime Data Anal.*, **15**, 120–146.
Peto, R., Collins, R. and Gray, R. (1995) Large-scale randomized evidence: large, simple trials and overviews of trials. *J. Clin. Epidem.*, **48**, 23–40.
Prevost, T. C., Abrams, K. R. and Jones, D. R. (2000) Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Statist. Med.*, **19**, 3359–3376.
Rosenbaum, P. R. (1991) A characterization of optimal designs for observational studies. *J. R. Statist. Soc.* B, **53**, 597–610.
Rosenbaum, P. R. (2002) *Observational Studies*, 2nd edn. New York: Springer.
Rosenbaum, P. R. and Rubin, D. B. (1983a) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.
Rosenbaum, P. R. and Rubin, D. B. (1983b) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Statist. Soc.* B, **45**, 212–218.
Rothwell, P. M. (2005a) External validity of randomised controlled trials: "To whom do the results of this trial apply?". *Lancet*, **365**, 82–93.
Rothwell, P. M. (2005b) Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*, **365**, 176–186.
Rubin, D. B. (1973) The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, **29**, 185–203.
Rubin, D. B. (1976) Inference and missing data (with discussion). *Biometrika*, **63**, 581–592.
Rubin, D. B. (1977) Assignment to treatment group on the basis of a covariate. *J. Educ. Statist.*, **2**, 1–26.
Rubin, D. B. (2001) Using propensity scores to help design observational studies: application to the tobacco litigation. *Hlth Serv. Outcms Res. Methodol.*, **2**, 169–188.
Rubin, D. B. (2008) For objective causal inference, design trumps analysis. *Ann. Appl. Statist.*, **2**, 808–840.

Rubin, D. B. and Thomas, N. (1992) Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika*, **79**, 797–809.
Rubin, D. B. and Thomas, N. (2000) Combining propensity score matching with additional adjustments for prognostic covariates. *J. Am. Statist. Ass.*, **95**, 573–585.
Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
Stuart, E. A. (2010) Matching methods for causal inference: a review and a look forward. *Statist. Sci.*, **25**, 1–21.
Stuart, E. A. and Green, K. M. (2008) Using full matching to estimate causal effects in non-experimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Devlpmntl Psychol.*, **44**, 395–406.
Sugai, G. and Horner, R. (2006) A promising approach for expanding and sustaining school-wide positive behavior support. *School Psychol. Rev.*, **35**, 245–259.
Sutton, A. J. and Higgins, J. P. (2008) Recent developments in meta-analysis. *Statist. Med.*, **27**, 625–650.
US Department of Education (2009) The impacts of regular Upward Bound on postsecondary outcomes seven to nine years after scheduled high school graduation. *Technical Report*. Office of Planning, Evaluation, and Policy Development, Policy and Program Studies Service, Washington DC.
US Department of Health and Human Services (2010) Head start impact study final report. *Technical Report*. Administration for Children and Families, Washington DC.
Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J. and Drazen, J. M. (2007) Statistics in medicine—reporting of subgroup analyses in clinical trials. *New Engl. J. Med.*, **357**, 2189–2194.
Weisberg, H., Hayden, V. and Pontes, V. (2009) Selection criteria and generalizability within the counterfactual framework: explaining the paradox of antidepressant-induced suicidality? *Clin. Trials*, **6**, 109–118.