Routledge
Taylor & Francis Group

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

Matching Methods for Selection of Participants for Follow-Up

Elizabeth A. Stuart & Nicholas S. Lalongo

To cite this article: Elizabeth A. Stuart & Nicholas S. Lalongo (2010) Matching Methods for Selection of Participants for Follow-Up, Multivariate Behavioral Research, 45:4, 746-765, DOI: 10.1080/00273171.2010.503544

To link to this article: https://doi.org/10.1080/00273171.2010.503544

Multivariate Behavioral Research, 45:746–765, 2010

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2010.503544

Psychology Press
Taylor & Francis Group

Matching Methods for Selection of Participants for Follow-Up

Elizabeth A. Stuart
Johns Hopkins Bloomberg School of Public Health

Nicholas S. Lalongo
Johns Hopkins School of Public Health, Department of Mental Health

This work examines ways to make the best use of limited resources when selecting individuals to follow up in a longitudinal study estimating causal effects. In the setting under consideration, covariate information is available for all individuals but outcomes have not yet been collected and may be expensive to gather, and thus only a subset of the comparison participants are followed. Expressions in Rubin and Thomas (1996) show the benefits that can be obtained, in terms of reduced bias and variance of the estimated treatment effect, of selecting comparison individuals well matched to those in the treated group compared with a random sample of comparison individuals. We primarily consider nonexperimental settings but also consider implications for randomized trials. The methods are illustrated using data from the Johns Hopkins University Baltimore Prevention Program, which included data collection from age 6 to young adulthood of participants in an evaluation of 2 early elementary-school-based universal prevention programs.

Longitudinal follow-up of participants is expensive, and nearly all longitudinal studies are faced with budget constraints. Often the resulting data collection effort does not have the resources to follow up all individuals fully. This will particularly be the case for studies collecting biologic data, which may be expensive to obtain or difficult to collect because of the intrusiveness of the data collection procedures. Standard practice has been to either spread the resources

Correspondence concerning this article should be addressed to Elizabeth Stuart, Department of Mental Health, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 624 North Broadway, Room 804, Baltimore, MD 21205. E-mail: estuart@jhsph.edu

out over the entire sample, which may lead to high non-response rates, or else to perhaps select a random sample of the participants to follow up. But are either of those the best strategy? Some recent work has investigated ways to maximize the information obtained on survey respondents by focusing resources on certain individuals (called "planned missingness"; Brown, Indurkhya, & Kellam, 2000; Graham, Taylor, Olchowski, & Cumsille, 2006), but work in this area is limited.

This article considers a setting where there is interest in estimating the effect of some "treatment" (or exposure), such as adolescent drug use or an early childhood intervention, using observational (nonexperimental) data. We imagine that we have a set of exposed individuals (e.g., drug users), as well as a larger set of comparison individuals (e.g., non-drug users), with baseline data (covariates) available on all individuals. The outcome data have not yet been collected. Resource constraints imply that not all individuals can be followed up. In particular, we follow up all of the exposed individuals (the drug users) but can afford to follow up only a subset of the comparison individuals (the non-drug users). This article addresses the question of how to select those comparison individuals for follow-up, in particular examining the benefits of selecting a set of comparison individuals most similar to the treated individuals versus a random set of comparison individuals. We in particular use the methods of propensity scores to select study participants. Although propensity scores are becoming an increasingly used tool for estimating causal effects in nonexperimental settings, their use in the design of longitudinal follow-up studies has been less emphasized.

This work builds on theoretical results regarding propensity scores in Rubin and Thomas (1992a, 1992b) and Rubin and Stuart (2006). Although these ideas have been around for many years and propensity scores are commonly used to select comparison participants in existing databases, they have not often been used in the data collection itself. This article extends the previous work in this area by focusing on practical implications of the approach, including guidance regarding when selecting matched versus random samples will be most useful (and when it can perhaps go wrong). Although many lessons from the general setting of nonequivalent control groups will also hold here, it is important to carefully consider the particular implications when selection is for data collection purposes.

Applications of the use of propensity scores in data collection include Reinisch, Sanders, Mortensen, and Rubin (1995), who examined the effects of prenatal exposure to phenobarbital on intellectual development. In that study, the covariates available included large numbers of prenatal records, there were approximately 100 exposed individuals and 8,000 potential comparison individuals, and the collection of outcomes (examinations and interviews when the individuals were approximately 20) was very expensive. Rather than selecting

a random subset of the 8,000 comparison individuals, those who looked the most similar to the exposed individuals were selected for follow-up. A second example is Hill, Rubin, and Thomas (2000), who used the same idea but in the context of a randomized experiment, estimating the effect of a school voucher program in New York City. In that case, control students (those who did not win the voucher lottery) were selected for follow-up based on their similarity to the treatment group (lottery winners).

The methods in this article are illustrated using fully simulated data as well as data from the Johns Hopkins Center for Prevention and Early Intervention Baltimore Prevention Program (BPP) first generation trials of classroom-based interventions (Kellam et al., 2008; Kellam, Rebok, Ialongo, & Mayer, 1994). Those studies were carried out in the 1980s and 1990s with all first graders in a set of Baltimore, MD public schools. The students have been followed since, including a current round of follow-up, when the students in the first trials are now in their early 30s. The first trial first involved randomization of a set of schools to treatment and comparison conditions, then randomization of classrooms within the intervention schools to intervention or control status, and then assignment of students to classrooms in a way that ensured balance of important covariates across classrooms. This unique design allows us to consider the implications of the methods for both the "internal" (same school) control students as well as the "external" (different school) control students, where the internal controls are somewhat more similar to the intervention groups than are the external controls. Thus, although this was originally a randomized trial, we use it to help learn about longitudinal follow-up more generally, for both nonexperimental studies and experiments.

This article proceeds as follows: the next section provides the theoretical basis for this work, summarizing results that show the reductions in bias in the treatment effect estimates that can be attained by using propensity score matching. This is followed by a simulation study that meets the distributional assumptions of those theoretical results, examining the settings under which matched samples yield particularly large benefits in relation to random samples. The following section uses a similar simulation approach but using the BPP data in order to investigate the method's performance in a more realistic setting. Finally, the article concludes with recommendations for practice and directions for future work.

THEORETICAL BASIS FOR USING PROPENSITY SCORES TO REDUCE BIAS

This work grows out of the literature on propensity scores, which generally are used to select participants for comparison when estimating causal effects. In

particular, we investigate the benefits of selecting a subset of the comparison group for follow-up: those who have propensity scores similar to those of the treated group. One of the key features of propensity scores, and one we take advantage of here, is that they are estimated using only covariate information: outcome information does not come into the estimation or use of propensity scores. Thus, they are particularly useful for settings like that considered here, where outcome data is not yet available, and in fact the propensity scores are used to help select the group for which we collect the outcome data.

Propensity scores were first developed in the 1980s by Rosenbaum and Rubin (1983). The propensity score is formally defined as the probability of receiving the treatment given the measured covariates. Propensity scores are typically estimated using logistic regression where the indicator of treatment receipt is regressed on the covariates; the predicted probabilities from that model are the propensity scores. Propensity scores can be used in matching, weighting, or subclassification (Stuart, in press). Because weighting or subclassification generally require outcome data on all individuals in the original sample, matching is the most appropriate method for the setting considered here. In fact, this setting of selecting participants for follow-up is what motivated some of the early work in matching methods (Rubin, 1973).

In particular, we propose the use of 1:1 nearest neighbor propensity score matching, where for each treated individual we select the comparison individual with the most similar propensity score. We do this "without replacement," which means that each comparison individual can be used as a match only once, and use a simple "greedy" algorithm. A greedy algorithm matches each treated individual one at a time, selecting from among those controls that have not yet been matched the control with the closest propensity score, without considering a global distance measure. A more sophisticated version might use an optimal algorithm, which would allow earlier matches to be broken if it would yield a lower global distance measure. Other possible refinements include selecting more than one comparison participant for each treated individual (k:1 matching) or combining propensity score matching with Mahalanobis metric matching on a few key covariates. Future work should investigate the pros and cons of alternative matching methods in this context. For example, k:1 matching might be useful if the control group is many times larger than the treated group. Other ways of using propensity scores, such as weighing or subclassification, are less appropriate here given the scenario of being able to collect outcome data on just a subset of the study sample. See Stuart (in press) for more discussion of this issue and that paper; Schafer and Kang (2008); and Shadish, Clark, and Steiner (2008) for more details on matching methods.

Once the matches are selected outcome data are collected on the full treatment group and their matched comparison participants, and analyses are done using those samples. One of the key properties of propensity scores implies that

matching on the propensity score can yield matched samples that have the same distributions of the full set of covariates, thus eliminating bias in the treatment effect estimate due to those covariates. In a series of papers, Rubin and Thomas (1992a, 1992b) and Rubin and Stuart (1996, 2006) formalized the benefits of selecting matched versus random samples for estimating treatment effects. One of the practical implications of that theoretical work is that it yields expressions for the bias reduction possible by selecting matched rather than random samples, using information only on the baseline characteristics (not any outcome data), which enables researchers to gauge the benefit they may obtain by selecting matched participants. The following sections summarize those results, which form the motivation for the 1:1 matching approach we consider.

Formal Setting

The formal setting we consider is one with two groups of individuals: N_t treated individuals and N_c control. We assume without loss of generality that $N_c > N_t$ ($R = \frac{N_c}{N_t}$) and that resources exist to follow up all treated individuals but only a subset of the controls of size $n_c = N_t$. This can be modified for settings where only a subset of the treatment group is followed up (Rubin & Stuart, 2006; Rubin & Thomas, 1992a, 1992b). There are p covariates X observed in both groups, where in the treated group $X \sim N(\mu_t, \Sigma_t)$ and in the control group $X \sim N(\mu_c, \Sigma_c)$. The discriminant between the groups is defined as $Z = (\mu_t - \mu_c)' \Sigma_c^{-1} X$ and is the linear combination of the covariates that leads to the largest difference between the groups. The propensity score can be thought of as a function of the discriminant because some function of the propensity score is often approximately linear in X (Rubin & Thomas, 1996). The results shown here are for matching using the estimated propensity scores; parallel results for situations where the true propensity scores are known are discussed in Rubin and Thomas (1992b). Interest is in estimating the effect of the treatment on some outcome, Y. For the theoretical results and analytic expressions we assume that Y is a linear function of the covariates X: $Y = \gamma X$. The estimand of interest is the average effect of the treatment on the treated, defined as $E(Y(1)|T = 1) - E(Y(0)|T = 1)$, where $Y(1)$ and $Y(0)$ are the potential outcomes under treatment and control, respectively, and T is an indicator of treatment assignment.

Two concepts that are important for the theoretical results are affinely invariant matching methods and ellipsoidally symmetric distributions. Affinely invariant matching methods are methods that result in the same set of matches after a linear transformation of the data (e.g., the same matches are obtained whether height is measured in feet or meters). Propensity score matching and Mahalanobis metric matching are two examples. Ellipsoidally symmetric co-

variate distributions are distributions such that a linear transformation of the variables yields a spherically symmetric distribution; they include the normal and t distributions (Dempster, 1969). For simplicity, we present the results for ellipsoidally symmetric distributions, but in fact the results hold in more general distributional settings, including conditionally ellipsoidal symmetric distributions, where the continuous covariates follow ellipsoidal distributions within categories defined by the categorical covariates (such as corresponds to a general location model) and mixtures of ellipsoidally symmetric distributions (Rubin & Stuart, 2006).

Two quantities that are important in predicting the amount of bias reduction possible when matching on the propensity score are (a) the initial covariate imbalance between the groups (the squared number of standard deviations between the covariate means of the treatment and control groups), calculated as the Mahalanobis distance: $B^2 = (\mu_t - \mu_c)'\Sigma_c^{-1}(\mu_t - \mu_c)$, and (b) the ratio of the variances of the discriminant in the treatment and control groups: $\sigma^2 = \frac{(\mu_t-\mu_c)\Sigma_c^{-1}\Sigma_t\Sigma_c^{-1}(\mu_t-\mu_c)}{(\mu_t-\mu_c)'\Sigma_c^{-1}(\mu_t-\mu_c)}$. Note that in a randomized trial $B^2 = 0$ and $\sigma^2 = 1$.

Bias and Variance Benefits of Selecting Matched Versus Random Samples

The first result shows that an affinely invariant matching method with covariates that follow ellipsoidally symmetric distributions is "equal percent bias reducing" (EPBR; Rubin & Thomas, 1992a, 1992b, 1996). An EPBR method reduces imbalance (differences) in all covariates by the same amount. The EPBR property is important because it ensures that reducing imbalance in one direction (e.g., the propensity score) will reduce imbalance in all directions, thus decreasing bias in the estimated treatment effect. Non-EPBR methods may increase imbalance in some covariates even while decreasing imbalance in others, which could lead to increased bias in the estimated treatment effect.

The EPBR property is shown more precisely here. With ellipsoidally symmetric distributions and affinely invariant matching methods, we can decompose Y into two parts: (a) its projection onto the discriminant Z and (b) the component W uncorrelated with Z. In other words, W is the portion of Y that is unaccounted for by the discriminant. Let ρ be the correlation between Y and Z. By construction, we know that the correlation of W and Z is 0. We then can decompose the effects of the matching into the effects on Z and on W. The key insight is that matching on Z cannot create imbalance in W because Z and W are uncorrelated. This implies that the reduction in bias of any estimated treatment effect on Y due to matching is proportional to the reduction in imbalance of Z. When estimating the treatment effect as a difference in

means of Y, the EPBR property can be expressed as follows (Rubin & Thomas, 1992a):

$$\frac{E(\overline{Y}_t - \overline{Y}_{mc})}{E(\overline{Y}_t - \overline{Y}_{rc})} = \frac{E(\overline{Z}_t - \overline{Z}_{mc})}{E(\overline{Z}_t - \overline{Z}_{rc})} = \tilde{g},$$

$$\tilde{g} = (1 - \theta^*_{max})_+,$$

where the expectations are over repeated samples from the population, the subscript t refers to the (full) treated group, rc refers to a random sample of size N_t from the control group, and mc refers to a matched sample of size N_t from the control group, obtained using a 1:1 propensity score match. θ^*_{max} is the maximum possible bias reduction and can be expressed as

$$\theta^*_{max} = \frac{\Omega(R/k)}{\{p(\sigma^2 + \frac{1}{R})N_t^{-1} + B^2\}^{1/2}}. \quad (1)$$

k is the number of controls selected for each treated participant (we assume $k = 1$ for simplicity), R is the ratio of the size of the (full) control group to the size of the (full) treated group (defined earlier), and $\Omega(R/k)$ is the expectation in the upper k/R tail of a standard normal distribution. Note that θ^*_{max} can be calculated using known quantities about the sample sizes and the distributions of X in the treated and control groups and thus can be used in advance to determine the amount of bias reduction that will be possible by selecting matched rather than random samples.

It is also possible to summarize the effects of the matching on the variance, although this quantity will differ for different Y because it depends on the correlation between Y and Z (ρ). The variance ratio is

$$\frac{var(\overline{Y}_t - \overline{Y}_{mc})}{var(\overline{Y}_t - \overline{Y}_{rc})} = \rho^2 \frac{var(\overline{Z}_t - \overline{Z}_{mc})}{var(\overline{Z}_t - \overline{Z}_{rc})} + (1 - \rho^2) \frac{var(\overline{W}_t - \overline{W}_{mc})}{var(\overline{W}_t - \overline{W}_{rc})}.$$

As with θ^*_{max}, this formula can be calculated using information known about the covariate distributions and sample sizes (see Rubin & Thomas, 1996).

These expressions provide guidance on what factors influence the bias reduction possible. B^2, the difference between groups on the covariates, is the factor that will affect the performance the most. Unfortunately, however, B^2 is not under the control of the researcher except inasmuch as it could inform the selection of control groups. Intuitively, the matching will also work best when R is relatively large because that implies there are many possible matches for each treated participant. However, the larger B^2 is the larger R will have to be

to get the same amount of bias reduction. For example, with an initial B^2 of 0.5, a ratio R of 2 will be sufficient to eliminate bias due to X. However, with an initial B^2 of 1.5, a ratio of 6 is required for the same bias reduction (Rubin & Thomas, 1996).

EVALUATION OF APPROACH: FULLY SIMULATED DATA

The theoretical results in the previous section indicate that there can be bias and variance benefits of selecting matched rather than random samples for long-term follow-up. This section examines the performance of the strategy of selecting matched versus random samples in practice using simulated data that meet the distributional requirements described in the previous section. The following section does similar evaluations using covariate distributions observed in a study of a behavioral program for first-grade students, which do not necessarily meet the distributional assumptions.

Simulation Setting

Interest is in comparing the bias and mean square error (MSE; as a function of both bias and variance) in the estimated treatment effect when selecting matched versus random samples. We use the general setting described earlier, with normally distributed covariates (which meet the condition of ellipsoidal symmetry). We consider non-linear outcome models, with the general form for Y being $Y_i(0) = Y_i(1) = e^{\frac{a}{\sqrt{p}}\sum_{j=1}^{p} X_i}$, where $Y_i(0)$ and $Y_i(1)$ are the potential outcomes under control and treatment, respectively, for individual i. Nonlinear models are used because linear regression adjustment would remove all bias if the true relationship is linear. We are interested in settings with mild to moderate nonlinearity, which can be hard to detect but that may yield considerable bias if there are covariate differences between the treatment and control groups.

We assume there is no treatment effect so that $Y_i(0) = Y_i(1)$ for all i; this has no loss of generality with respect to constant treatment effects across individuals. Future work will investigate settings with heterogeneous treatment effects. We consider a setting with a relatively small population, with 200 treated and 200 control individuals in the full population. For each simulation we randomly draw 75 treated participants, so $N_t = n_t = 75$, and we use either $N_c = 100$ or $N_c = 150$ controls, depending on the simulation setting. Of those possible controls, $n_c = n_t = 75$ are picked either randomly or using propensity score matching. The simulations vary:

1. The number of covariates: $p = 1, 3, 5$. All covariates are generated as Normal(μ,1).
2. The initial number of standard deviations difference between the treated and control groups on X: $\delta_t = 0, .5, 1$ ($B^2 = \delta_t^2 * p$).
3. The ratio of the size of the control group to the size of the treatment group (N_c/N_t): $R = 4/3, 2$.
4. The amount of nonlinearity in the relationship between the covariates and outcome: $a = .1, .5, 1$.

The values of a considered correspond to linear R^2 values of approximately 0.99, 0.85, and 0.6. Although primary interest here is in nonexperimental studies, the settings with $\delta_t = 0$ correspond to what would be expected in a randomized experiment or a nonexperimental study with very well balanced covariates starting out. We also examined settings with $a = 1.5$; the results were similar to those presented here in terms of the relative performance of matched versus random samples but are not included because of the large absolute size of the bias and MSE of those estimates. Similarly, larger values of p, the number of covariates, such as $p = 10$, led to large absolute values of the bias and MSE for both approaches, dominating the summaries and making it harder to clearly see differences between the methods. We discuss this issue further later; the values of R used here are relatively small, and with higher values of R it would likely be very possible to match on a larger number of covariates.

At each simulation setting we created 500 simulated data sets, where for each simulated data set we calculated the bias and MSE in the estimated treatment effect calculated in four different ways: matched samples; random samples; and for each, differences in means and regression adjustment controlling for the covariates X. The matched samples were obtained using a 1:1 nearest neighbor propensity score match (Stuart, in press) where the propensity scores were estimated using logistic regression with each of the covariates X included as predictors. The comparison of matched and random samples is quantified by the ratio of average bias (or MSE) in the estimated treatment effect in the matched samples compared with the random samples. A ratio less than 1 implies that the matched samples yielded lower bias (or MSE) in the estimated treatment effect, on average, compared with the random samples.

The matching was carried out using the MatchIt package (Ho, Imai, King, & Stuart, in press) for the R software program (R Development Core Team, 2008).[1] Additional software for performing propensity score matching is described in

[1]Implementing 1:1 greedy nearest neighbor matching without replacement in MatchIt requires just one line of code: m.out <− matchit(treat ∼ x1 + x2 + x3, data=dta). See the MatchIt documentation for details on how to modify that code for k:1 matching, matching with replacement, or optimal rather than greedy matching.

(Stuart, in press) and online at http://www.biostat.jhsph.edu/∼estuart/propensityscoresoftware.html

Results

Table 1 summarizes those results by averaging across simulation settings to provide summaries for each factor considered. This averaging across other factors is justified by the fact that there are no factor-by-factor interactions that are significant in analysis of variance (ANOVAs) of the bias and MSE on the factors and all interactions of factors. Examining only the main effects of each factor thus gives a meaningful summary of the results. In the left side of Table 1 we see that across settings, when estimating the treatment effect using a simple difference in means of the outcome, the ratios are always less than 1, indicating that the matched samples yielded lower bias and MSE than the random samples. The only exceptions to this occurred in a few particular simulation draws when $\delta_t = 0$ (i.e., $B^2 = 0$), which corresponds to a situation where the original treated and control groups are actually only randomly different from one another. However, in that setting even when the random samples yielded lower bias in the treatment effect, the difference was minimal (e.g., a ratio of average biases of approximately 1.03), and the MSE was always lower in the matched samples. In addition, in such a setting the absolute bias of each of the methods was very small.

When the effect estimates are obtained using regression adjustment that controls for the baseline covariates X, the bias and MSE for both methods (matched and random samples) is decreased substantially: The absolute values of the bias and MSE are lower for both matched and random samples compared with the difference in mean results. However, the matching method generally still shows better performance than random samples, except in cases where both methods essentially provide unbiased estimates of the treatment effect. In some cases the difference is substantial. For example, when the control pool is relatively large relative to the size of the treatment group ($R = 2$), the ratio of bias in matched versus random samples is 0.42.

The simulations also provide insight into what scenarios are particularly problematic and in which settings matching can yield the most improvement. In general, the largest bias without matching is found when there are large initial differences in the covariates X (large values of δ_t), a high degree of nonlinearity in the relationship between the covariates X and outcome Y (large a), or a large number of covariates (large p). Matching helps the most (in terms of the percentage reduction in bias in the estimated treatment effect) under the settings under which it is easiest to find good matches: a relatively small amount of initial bias in X (small δ_t), many more control than treated units (large R), or a small number of covariates (small p). The relative size of the control pool is

TABLE 1
Summary of Relative Performance of Matched Versus Random Samples

	Difference in Means						*Regression Adjustment*					
	Bias			*MSE*			*Bias*			*MSE*		
	Matched	*Random*	*Ratio*	*Matched*	*Random*	*Ratio*	*Matched*	*Random*	*Ratio*	*Matched*	*Random*	*Ratio*
Overall	1.17	1.37	0.85	7.86	9.28	0.85	−0.16	−0.24	0.68	1.03	1.03	1.00
$R = 4/3$	1.25	1.37	0.91	8.42	9.28	0.91	−0.23	−0.24	0.95	1.27	1.02	1.24
$= 2$	1.09	1.37	0.79	7.30	9.28	0.79	−0.10	−0.24	0.42	0.78	1.04	0.75
$p = 1$	0.46	0.62	0.74	0.82	1.32	0.62	0.16	0.05	3.03	0.09	0.03	2.82
$= 3$	1.12	1.34	0.83	5.50	6.95	0.79	−0.09	−0.17	0.53	0.22	0.26	0.84
$= 5$	1.93	2.16	0.90	17.27	19.55	0.88	−0.56	−0.60	0.93	2.78	2.80	0.99
$\delta_t = 0$	0.02	0.04	0.56	0.01	0.02	0.59	0.02	0.03	0.68	0.01	0.02	0.69
$= 0.5$	0.69	0.96	0.72	1.19	2.01	0.59	0.10	−0.03	−3.79	0.05	0.05	1.20
$= 1$	2.80	3.12	0.90	22.38	25.79	0.87	−0.61	−0.72	0.84	3.01	3.03	1.00
$a = 0.1$	0.06	0.09	0.66	0.01	0.02	0.53	0.00	0.00	6.63	0.00	0.00	2.03
$= 0.5$	0.56	0.72	0.77	0.75	1.08	0.70	−0.02	−0.03	0.51	0.02	0.01	1.33
$= 1$	2.89	3.31	0.87	22.82	26.74	0.85	−0.47	−0.69	0.69	3.06	3.08	1.00

Note. Numbers shown are the ratio of the average bias or mean square error (MSE) between samples selected using propensity score matching and samples selected randomly. Results from simulated data with ellipsoidally symmetric covariate distributions.

particularly crucial: we see only moderate benefit of matching when $R = 4/3$ but substantial benefit when $R = 2$. Larger ratios would lead to even greater improvement as it would facilitate the selection of even better matches for the treatment group members. There is some residual bias in our setting even when using matching because the control pools are not large enough to obtain exact matches and remove all bias. This is also why the regression adjustment on top of matching yields lower bias—it adjusts for the small covariate differences that remain after matching. In addition, the expression for θ^*_{max} given in Equation (1) was quite accurate as also found in Rubin and Thomas (1996). Because other work (Rubin & Thomas, 1996; Stuart, 2004) shows detailed results on the quality of the expressions, we do not present those results here.

EVALUATION OF APPROACH: SIMULATIONS BASED ON BPP DATA

We next consider data from an evaluation of two school-based interventions to improve academic achievement and reduce aggression and problem behavior: Mastery Learning (ML) and the Good Behavior Game (GBG). Beginning in 1985, two cohorts of first graders within 19 elementary schools in an urban mid-Atlantic region in the United States were enrolled in the study ($N = 2{,}311$), and have been followed since. Data collection activities included yearly surveys and teacher reports of behavior through elementary school and periodic follow-up going into the sample participants' late 20s and early 30s; the most recent data collection is currently ongoing. The design involved both school-level and classroom-level randomization. First, schools were randomized to one of the two programs (ML or GBG) or to the control condition. Then, first-grade classrooms within each of the program schools were randomized to receive the program or serve as a control. Finally, students were assigned to classrooms in a balanced manner that led to similar student characteristics across the classrooms. This led to two types of controls: internal controls within the program schools (the classrooms within those schools that were assigned to the control condition) and external controls in the schools that were randomized to be control schools. Although the trial itself was done with two cohorts of children, for simplicity and illustration we combine the two cohorts into one analysis.

Because evaluating the procedures requires knowing the true treatment effect, we use the observed covariate data, combined with simulated outcomes, with known treatment effects. Tables 2 and 3 show the covariates and their means used in this investigation for the ML and GBG samples, respectively. Because the purpose of this article is purely illustrative, we restricted our analyses to individuals with fully observed data on the covariates of interest, and thus the numbers in Tables 2 and 3 differ somewhat from other studies using this data.

TABLE 2
Covariates From Baltimore Prevention Program Study: Mastery Learning

		Internal Controls		*External Controls*	
Covariate	*ML*	*M*	*p Value*	*M*	*p Value*
Male	48%	53%	.17	52%	.23
African American	66%	62%	.23	65%	.70
Eligible for free or reduced lunch (FRPL)	54%	52%	.65	45%	.00
Standardized test score (Test)	271	275	.21	266	.05
Aggression (Aggress)	1.71	1.68	.65	1.93	.00
Conduct problems (Conduct)	2.03	1.90	.39	1.68	.01
N	444	351		515	

Note. *p* values shown are for difference between internal and external controls, respectively, and the Mastery Learning (ML) treatment group. *p* values from χ^2 test for binary covariates, *t* test for continuous covariates. All covariates measured in fall of first grade.

In Tables 2 and 3 we see that the internal controls are generally more similar to the intervention groups than are the external controls, at least on the covariates considered here.

Table 4 illustrates the use of Equation (1) to estimate the amount of bias reduction attainable using these samples. The most important difference between the settings is the initial covariate imbalance between the groups (B^2) and R, the total size of the control pool relative to the number of control matches to be picked. Table 4 summarizes these quantities for the BPP data, considering both the GBG and ML groups and the internal and external controls.

TABLE 3
Covariates From Baltimore Prevention Program Study: Good Behavior Game

		Internal Controls		*External Controls*	
Covariate	*GBG*	*M*	*p Value*	*M*	*p Value*
Male	50%	47%	.37	52%	.72
African American	78%	69%	.01	65%	.00
Eligible for free or reduced lunch (FRPL)	60%	64%	.25	45%	.00
Standardized test score (Test)	267	266	.82	266	.70
Aggression (Aggress)	1.97	1.93	.10	1.93	.52
Conduct problems (Conduct)	1.59	1.81	.25	1.68	.57
N	414	290		515	

Note. *p* values shown are for difference between internal and external controls, respectively, and the Good Behavior Game (GBG) treatment group. *p* values from χ^2 test for binary covariates, *t* test for continuous covariates. All covariates measured in fall of first grade.

TABLE 4
Maximum Amount of Bias Reduction Possible When Selecting Matched Samples Using Estimated Propensity Scores in the Baltimore Prevention Program Data

Comparison	B^2	R_c	Θ^*_{max}
GBG vs. internal controls	0.08	1.16	1.22
GBG vs. external controls	0.13	2.06	2.27
ML vs. internal controls	0.03	1.40	1.92
ML vs. external controls	0.17	2.06	2.17

Note. GBG = Good Behavior Game; ML = Mastery Learning.

The fact that all of the Θ^*_{max} values in Table 4 are greater than 1 indicates that 100% bias reduction is possible in each of the four comparisons by selecting matched samples. This of course is a good situation to be in and is not surprising given that the BPP study was a randomized trial. Larger values of B^2 would lead to Θ^*_{max} values less than 1, indicating that some, but not full, bias reduction is possible. The values of B^2 in Table 4 confirm that the internal controls are somewhat more similar to the intervention groups compared with the external controls, as also seen in the individual characteristics in Tables 2 and 3. We also see that the bias reduction possible is larger (Θ^*_{max} is larger) for the external controls compared with the internal; the slightly larger initial covariate imbalance (B^2) is counterbalanced by the larger number of possible controls (represented by R), meaning that more bias reduction is possible.

Because the covariates do not fully follow ellipsoidally symmetric distributions, the results in Table 4 are approximations and should be viewed as providing general guidance. We now turn to simulations that directly address the bias and MSE of average treatment effects estimated using matched and random samples parallel to the simulations presented earlier. We consider the full population all students in the study. For example, for the ML comparisons, $N_t = 444$, $N_c = 515$ for the external comparisons and $N_c = 351$ for the internal comparisons. To simulate a study that contained only a subsample of the population, at each iteration of the simulation, we drew a random subsample from the treated group of size $n_t = 250$ and matched and random subsamples of size $n_c = n_t = 250$ from the appropriate control group (internal or external). As described earlier, we used a 1:1 nearest neighbor greedy matching algorithm to select the matched samples. The propensity score model used to select the matches was estimated using a logistic regression with each of the covariates listed in Tables 2 and 3 as predictors. We performed 100,000 simulations using the internal controls and 100,000 using the external controls and repeated this whole procedure twice, once for the GBG and once for ML. As in the simulations described earlier, we

consider the performance of the approaches when effects are estimated using a simple difference in means as well as using regression adjustment.

We examine two outcomes, one with residual error added in and one without residual error, to more clearly investigate the effects of the matching on bias (as in Rubin & Thomas, 1996, and Stuart & Rubin, 2008). Again the outcome model was chosen to be nonlinear in the covariates because regression adjustment would provide unbiased estimates of the treatment effects if the linear model was correct and was selected to reflect the distribution and relationships observed in the data between the covariates and a sample outcome, achievement test scores in fourth grade. The outcome models used were

$$\begin{aligned} Y_1 &= 150 + 15 * \text{Male} - 16 * \text{AfricanAmerican} - 8 * \text{FRPL} + \\ &\quad .75 * \text{Test} - 2 * \text{Aggress} - 3 * \text{Conduct} - \\ &\quad .0006 * \text{Test}^2 - 1.8 * \text{Aggress}^2 \\ Y_2 &= Y_1 + N(0, 100). \end{aligned}$$

Tables 5 and 6 provide a summary of the results, showing the bias and MSE for each of the outcomes considered. In general the results are quite consistent with the results using the fully simulated data. Across nearly all conditions the matched samples yielded lower bias and MSE compared with the random samples. When estimating the treatment effect using a difference in means, the benefit of matched samples was substantial. For the ML intervention, the bias and MSE ratios range from 0.05 to 0.27, meaning that the matched samples yielded treatment effect estimates with at most one quarter the bias and MSE of results from random samples. For the GBG the ratios ranged from 0.18 to 0.69. When the treatment effect is estimated using regression adjustment, the matching method still generally shows better performance than random samples, but there is less of a difference between the approaches. With regression adjustment both approaches (matched and random samples) yielded relatively low bias and MSE, and the ratios are closer to 1 than they were for the difference in mean estimates. Less benefit of matching is also found when using the internal controls. This is as expected given the closer similarity of the internal controls to the treated group compared with the external controls. However, the matching makes a large difference when using the external controls, especially for the first outcome, with ratios of approximately 0.4 for ML and 0.6 for the GBG when treatment effects are estimated using regression adjustment, and even smaller ratios when effects are estimated using a difference in means.

DISCUSSION

The efficient design of longitudinal studies is an important topic that has to this point received relatively little research attention. The standard approach currently is to follow up the full sample of individuals. However, resources are often limited and it may not be cost-effective (or possible) to follow up the full original study sample. In fact, fairly often the baseline data are available, or relatively inexpensive to obtain, but the outcome data are expensive. This article has investigated a possible alternative for when

TABLE 5
Results From Baltimore Prevention Program Simulations: Mastery Learning

	Difference in Means						*Regression Adjustment*					
	Bias			*MSE*			*Bias*			*MSE*		
	Matched	*Random*	*Ratio*	*Matched*	*Random*	*Ratio*	*Matched*	*Random*	*Ratio*	*Matched*	*Random*	*Ratio*
						Internal Controls						
Y_1	0.32	−1.56	0.21	1.07	6.02	0.18	0.70	0.72	0.97	0.51	0.54	0.94
Y_2	0.14	−1.65	0.08	1.46	6.85	0.21	0.49	0.57	0.86	0.80	0.90	0.89
						External Controls						
Y_1	0.17	3.53	0.05	1.24	16.63	0.07	0.07	0.20	0.35	0.03	0.07	0.43
Y_2	1.10	4.15	0.27	3.00	22.03	0.14	1.00	1.05	0.95	1.61	1.81	0.89

Note. MSE = mean square error.

TABLE 6
Results From Baltimore Prevention Program Simulations: Good Behavior Game

	Difference in Means						*Regression Adjustment*					
	Bias			*MSE*			*Bias*			*MSE*		
	Matched	*Random*	*Ratio*	*Matched*	*Random*	*Ratio*	*Matched*	*Random*	*Ratio*	*Matched*	*Random*	*Ratio*
						Internal Controls						
Y_1	−1.09	−1.70	0.64	2.49	6.85	0.38	−0.76	−0.77	0.99	0.62	0.63	0.99
Y_2	−1.09	−1.59	0.69	3.00	6.85	0.44	−0.85	−0.84	1.01	1.16	1.17	1.00
						External Controls						
Y_1	−0.90	−2.98	0.30	2.45	13.74	0.18	−0.26	−0.38	0.68	0.11	0.19	0.57
Y_2	−0.98	−3.09	0.32	3.33	15.24	0.22	−0.33	−0.38	0.85	0.79	0.89	0.89

Note. MSE = mean square error.

resources are limited and it is not possible to follow up all study participants. We have shown that selecting for follow-up comparison participants who are well matched to the treated individuals (compared with selecting comparison participants randomly) can yield improved bias and MSE of an estimated treatment effect.

In particular, the results provide guidance for the design of longitudinal studies. They show that in nonexperimental studies it can be very beneficial to select matched rather than random subsamples for follow-up. This is especially true when the effect estimate will be estimated using a simple difference in means but is also true when regression adjustment will be used to control for small covariate differences between the groups. This is the same situation as in nonexperimental studies more generally when outcome data may already be available. This work complements that research by highlighting the usefulness of matching methods such as propensity scores in the data collection phase as well. In randomized experiments there is less benefit to selecting matched samples, although it will generally not be harmful; matched and random samples perform similarly with respect to bias and MSE, with both methods yielding unbiased treatment effect estimates.

There are a number of limitations of this work and complications that have not yet been addressed. One is the issue of clustering. Many randomized trials, including the BPP, involve either group randomization or individuals who are individually randomized but grouped into clusters, such as in interventions that involve group sessions. The methods currently ignore that clustering. Determining how to incorporate the clustering into the selection of participants for follow-up will be an important area for future research. One possibility will be methods such as those in Stuart and Rubin (2008), who described a procedure for selecting matches from across multiple clusters. This would also provide another way to take advantage of settings with multiple control groups, such as the internal and external controls in the BPP. Another issue is that this work assumes that there are more control individuals than treated and thus it is possible to subset the controls for follow-up. Implications for settings with fewer control than treated, or similar size groups, should be considered. A final limitation is the limited nature of the simulations, in particular the assumption of a constant treatment effect and the relatively small number of covariates. Future work should expand these simulations to more realistic settings.

With respect to the number of covariates, the results show that matching is most helpful with very small numbers of covariates (five or fewer). With more covariates both approaches have much higher bias and MSE, and the distinction between methods is less clear. Matching may thus be most useful for selecting participants for follow-up when there is a small number of particularly prognostic covariates, such as pretreatment measures of the outcome. However, our settings had a relatively small ratio of control to treated participants (either 4/3 or 2), and when the ratio of control to treated participants is larger it will likely be possible (and beneficial) to match on larger numbers of covariates. The initial covariate imbalance between the groups will also influence on how many covariates good matches can be obtained.

This work also points toward the need for further research in the optimal design of longitudinal studies. One question relates to which covariates should be prioritized in the matching process: those related to treatment receipt or to the outcome, or both? For example, are these methods worth pursuing if the available covariates are not very

predictive of the outcome of interest? This may also have relevance for propensity score methods more generally and is a topic of ongoing research in the propensity score literature (Brookhart et al., 2006; Shadish et al., 2008), although as discussed earlier attention would need to be paid to whether the considerations are different in this scenario of selecting participants for follow-up compared with the standard use of propensity scores in nonexperimental studies. Another direction for future work is formalization of the cost trade-offs and recommendations for the size of the samples. In a randomized experiment, if it is determined that a random sample will suffice, then standard power analyses can be used to determine the sample size needed to detect the expected intervention effects. However, when a matched sample will be selected, as is particularly appropriate for nonrandomized studies, the sample may be able to be smaller than would be required for a random sample because of the improved performance shown earlier. To formalize this trade-off one could calculate the additional number of participants that would be required to follow up, if the sample is selected randomly instead of by matching, to obtain an estimated effect with similar bias and MSE.

In summary, this work has illustrated the potential bias and MSE benefits that may be obtained by selecting matched versus random samples for long-term follow-up. We hope that in part this work serves to increase discussion of strategies for the efficient design of longitudinal studies, helping researchers think through ways to make the most of the resources that they have.

ACKNOWLEDGMENTS

This work was supported by National Institute of Mental Health (NIMH) grants MH083846 (PI: Stuart) and MH066247 (PI: Lalongo). Thanks to Amy Goldstein at NIMH for helping to motivate this work and to Hendricks Brown and the Prevention Science Methodology Group (NIMH Grant R01MH040859) for helpful discussions.

REFERENCES

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163,* 1149–1156.

Brown, C. H., Indurkhya, A., & Kellam, S. G. (2000). Power calculations for data missing by design: Applications to a follow-up study of lead exposure and attention. *Journal of the American Statistical Association*, *95,* 383–395.

Dempster, A. P. (1969). *Continuous multivariate analysis*. Reading, MA: Addison-Wesley.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11,* 323–343.

Hill, J., Rubin, D. B., & Thomas, N. (2000). The design of the New York School Choice Scholarship Program evaluation. In L. Bickman (Ed.), *Research designs: Inspired by the work of Donald Campbell* (chap. 7, pp. 155–180). Thousand Oaks, CA: Sage.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (in press). Matchit: Nonparametric preprocessing for parametric causal inference. Forthcoming in *Journal of Statistical Software*. Available from http://gking.harvard.edu/matchit/

Kellam, S. G., Brown, C. H., Poduska, J. M., Ialongo, N. S., Wang, W., Toyinbo, P., et al. (2008). Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. *Drug and Alcohol Dependence*, *95*(Suppl. 1), S5–S28.

Kellam, S., Rebok, G., Ialongo, N., & Mayer, L. (1994). The course and malleability of aggressive behavior from early first grade into middle school: Results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *35,* 259–281.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Available from http://www.cran.r-project.org. Vienna: R Foundation for Statistical Computing.

Reinisch, J., Sanders, S., Mortensen, E., & Rubin, D. B. (1995). In utero exposure to phenobarbital and intelligence deficits in adult men. *Journal of the American Medical Association*, *274*, 1518–1525.

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

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., & Stuart, E. A. (2006). Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *The Annals of Statistics*, *34,* 1814–1826.

Rubin, D. B., & Thomas, N. (1992a). Affinely invariant matching methods with ellipsoidal distributions. *Annals of Statistics*, *20*, 1079–1093.

Rubin, D. B., & Thomas, N. (1992b). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika*, *79*, 797–809.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores, relating theory to practice. *Biometrics*, *52*, 249–264.

Schafer, J. L., & Kang, J. D. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated case study. *Psychological Methods*, *13,* 279–313.

Shadish, W. R., Clark, M., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103,* 1334–1344.

Stuart, E. A. (2004). *Matching methods for estimating causal effects using multiple control groups*. Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Stuart, E. A. (in press). Matching methods for causal inference: A review and a look forward. Forthcoming in *Statistical Science*.

Stuart, E. A., & Rubin, D. B. (2008). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics*, *33,* 279–306.