Journal of Educational and Behavioral Statistics
Month XXXX, Vol. XX, No. X, pp. X–XX
DOI: 10.3102/1076998607306078
© AERA and ASA. http://jebs.aera.net

Matching With Multiple Control Groups With Adjustment for Group Differences

Elizabeth A. Stuart
Johns Hopkins Bloomberg School of Public Health

Donald B. Rubin
Harvard University

When estimating causal effects from observational data, it is desirable to approximate a randomized experiment as closely as possible. This goal can often be achieved by choosing a subsample from the original control group that matches the treatment group on the distribution of the observed covariates. However, sometimes the original control group cannot provide adequate matches for the treatment group. This article presents a method to obtain matches from multiple control groups. In addition to adjusting for differences in observed covariates between the groups, the method adjusts for a group effect that distinguishes between the control groups. This group effect captures the additional otherwise unobserved differences between the control groups, beyond that accounted for by the observed covariates. The method is illustrated and evaluated using data from an evaluation of a school drop-out prevention program that uses matches from both local and nonlocal control groups.

Keywords: *causal inference; historical data; nonrandomized data; observational study; propensity scores*

1. Introduction

1.1 Matched Sampling in Observational Studies

Matching methods, used in the context of causal inference to select groups of treated and control units with similar values of background covariates, have been receiving increasing attention over the past few decades in fields such as statistics (e.g., Rosenbaum, 2002; Rubin, 1973a, 1973b), economics (e.g.,

The work for this article was done while Elizabeth Stuart was a graduate student at Harvard University Department of Statistics and was partially supported by a National Science Foundation Graduate Research Fellowship. Elizabeth Stuart also acknowledges the support of the National Institute of Mental Health and the National Institute of Drug Abuse (MH066247). We particularly thank Neal Thomas and Gary King for helpful comments and Roberto Agodini of Mathematica Policy Research, Inc. for supplying us with the illustrative data.

Dehejia & Wahba, 1999; Imbens, 2004; J. Smith & Todd, 2005), political science (e.g., Ho, Imai, King, & Stuart, 2007), sociology (e.g., Morgan & Harding, 2006; H. Smith, 1997), and medicine (e.g., Christakis & Iwashyna, 2003). A very recent summary of three decades of work in matching is Rubin (2006). The general scenario involves selecting well-matched subsets of units from the original treatment and control groups to reduce bias due to those covariates when estimating the causal effect of treatment versus control.

However, in some settings, there may be interest in combining information from multiple control groups, for example, randomized experiments in which it is difficult or expensive to form a large control group but there are reliable historical patient data or a national disease registry to supplement the randomized controls or settings where the original control group does not contain enough units who look similar on observed covariates to those in the treatment group, as in the motivating example of this article, described in Section 1.4. When there are multiple control groups available, it may be wise to use good matches from each of these groups while simultaneously accounting for potential differences between the groups that are not accounted for by the observed covariates. For example, when using historical data to supplement a current randomized clinical trial, researchers may want to account for differences due to temporal changes not reflected in the matching variables. Here we consider situations with two control groups and find well-matched units from both groups to estimate and thereby adjust for simple differences between the control groups not reflected in the observed matching variables. Specifically, because values of the potential outcomes under control, $Y(0)$, are observed in both control groups, the difference in $Y(0)$ values between well-matched units from the two control groups can be used to try to adjust for differences between these groups when analyzing the treated and matched control data.

The article proceeds as follows. The general framework of causal inference is reviewed in Section 1.2, followed by a summary of previous uses of multiple control groups in Section 1.3 and a description of the motivating example, the evaluation of a school drop-out prevention program (the School Dropout Demonstration Assistance Program; SDDAP) in Section 1.4. Section 2 describes a matching method for use with two control groups, including an approximation for the optimal number of matches to obtain from each control group. Section 3 provides a description of the matching adjustment procedure. Sections 4 and 5 present evaluations of the method, first using simulated data and then using some SDDAP data. Finally, Section 6 concludes.

1.2. Conceptual Framework

We consider a study with one group that received the treatment of interest and two (or more) control groups that did not. A collection of covariates, X, is observed in all groups. The goal is to choose subsamples from the original control groups that match the treatment group on X, thereby reducing bias in the estimated treatment effect due to those covariates. We assume that interest focuses on estimating the average treatment effect in the full treatment group, and thus the matching is allowed to discard "irrelevant" members of the control groups but the full treatment group is retained. We also assume that the treatment assignment mechanism into the treatment group versus the original control group is ignorable given X (Rubin, 1978).

Causal effects inherently involve a comparison of potential outcomes under different treatments on a common set of units. For each individual unit i, we observe either $Y_i(1)$, the potential outcome under treatment, or $Y_i(0)$, the potential outcome under control, depending on treatment assignment. The estimand is the average treatment effect for the treated: $\tau = \frac{1}{n_t}\sum_{i \in \mathcal{T}}(Y_i(1) - Y_i(0))$, where $\mathcal{T}$ represents the treatment group and n_t is the number of treated units. Of course we never observe $Y_i(0)$ for the treatment group, and thus to estimate τ we effectively need to impute each treated unit's potential outcome under control, $Y_i(0)$. To do so, we seek control units who look similar to the treated units on all covariates, thereby effectively modeling the potential outcomes for the treated if they were exposed to the control. This sort of matching is often done using the propensity score (Rosenbaum & Rubin, 1983), which is the probability of receiving the treatment given the observed covariates.

By the definition of ignorability, the original control group (Control Group 1) is the primary control group in the sense that the only collection of covariates that need to be controlled when comparing to the treatment group is assumed to be X, but this is not true for the supplemental control group (Control Group 2). For example, the treatment and primary control group may be from the same geographical region, whereas the supplemental control group is from another region. Thus, Control Group 1 exactly matches the treatment group on area-level covariates but may not have good overlap with that group on the individual-level covariates X. In contrast, Control Group 2 may have good overlap with the treatment group on the individual-level covariates but is not from the same geographic area as the treatment group. Our objective is to form a single set of matched control units, with some matches chosen from each of the two potential control groups, to benefit from having both control groups.

The situation we consider is illustrated in Figure 1 with univariate X, where there is limited overlap between the treatment group and Control Group 1 (two standard deviations difference between the means in this hypothetical example). The region of X between the two vertical lines at $X = 0$ and $X = 4$ indicates values of X where there is reasonable overlap between the treatment group and Control Group 1 and between these groups and Control Group 2. For individuals in the treatment group with X values greater than about 0, there is a good match from Control Group 1. However, for individuals in the treatment group with X values less than about 0, there are few or no appropriate matches from Control Group 1. Those individuals will instead be matched to individuals from Control Group 2, which has good overlap with the treatment group over the full X

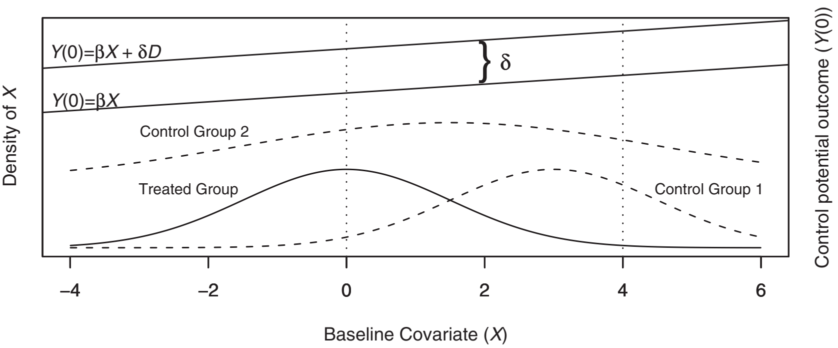

FIGURE 1. *Adjustment scenario.*

Note: Density plots illustrate the limited overlap between the treatment group and Control Group 1 on the baseline covariate X as well as the substantial overlap between Control Group 2 and both Control Group 1 and the treatment group. Regression lines indicate the estimation of δ, discussed in Section 2, using well-matched controls from Control Groups 1 and 2.

distribution. However, by assumption, the difference between the treatment group and Control Group 1 is captured by X, but the difference between Control Group 2 and the other two groups is not fully captured by X. If we are willing to restrict estimation of the treatment effect to the space of X where there is sufficient overlap between the treatment group and Control Group 1, for example, above about 0 or 1 in Figure 1, then we could simply use the few matches from Control Group 1 that match units in the treatment group and discard treated units without good matches from Control Group 1. However, in the setting of this article, we are interested in estimating the treatment effect for the full range of X values in the treatment group, and we are not willing to rely on extrapolation of the functional form of the model for $Y(0)$ given X in Control Group 1 to estimate the treatment effect for treated units with values of X outside the range of Control Group 1.

1.3. Previous Uses of Multiple Control Groups

There have been some previous uses of multiple control groups, generally in the context of testing for hidden bias. In particular, Campbell (1969) and Rosenbaum (1987) discussed using multiple control groups to estimate bounds on treatment effects or to corroborate results by assessing whether results obtained using multiple control groups are as expected given additional available information. For example, Campbell advocated using two groups that “bracket” the treatment group on some unobserved variable of concern, also known as “control by systematic variation.” Specifically, in a study of a supplemental educational program in which students must be eligible to participate but self-select into the program, a researcher could compare the treatment group both with a group of matched controls who were eligible but chose not to participate (presumed to have lower motivation than the students in the program) and a group of matched controls who were ineligible because of higher performance (presumed to have higher motivation than the students in the program). Lu and Rosenbaum (2004) specified an optimal matching algorithm to form pairwise matches from one treatment and two control groups, thus allowing three pairwise comparisons between the three groups. They illustrated the method using data from Card and Krueger's (1994) analysis of the effect of an increase in the minimum wage law in New Jersey, comparing the restaurants affected by the law change in New Jersey with two groups of control restaurants: restaurants in Pennsylvania that were unaffected because the law did not apply to them and restaurants in New Jersey that were unaffected because their starting wage was already above the new standard.

Multiple control groups have often been used in medicine, particularly through the use of historical controls to supplement information from a randomized or contemporaneous control group. Baker and Lindeman (1994, 2001) used multiple control groups to examine the effect of the availability of epidural anesthetic on the rate of Cesarean sections. Dempster, Selwyn, and Weeks (1983) treated a set of historical control groups as arising from a superpopulation and used Bayesian methods to pool the data. Using untreated historical patients to provide information on long-term trends in the outcome is illustrated in Shen and Fleming (1999) and Rubin, Cook, and Stuart (2003).

Rosenbaum (1987, 2002) provided a thorough examination of the use of multiple control groups, including formal discussion of the possible benefits of using two control groups, but he focused on the use of multiple control groups to test for hidden bias. Rosenbaum stressed that the value of a second source of controls depends critically on supplementary information that is available regarding unobserved biases that may exist. In particular, when some of this supplementary information is available, a second source of control units can be used to test the assumption of strongly ignorable treatment assignment to the three groups (Rosenbaum & Rubin, 1983), which states that treatment assignment is independent of the potential outcomes given the covariates. Essentially, if after adjusting for the observed covariates the two control groups differ with respect to the potential outcome under control, then the treatment assignment cannot be strongly ignorable, and at least one of the control groups is not comparable to the treatment group. We extend that approach by using the two control groups together in one analysis to adjust explicitly for the "hidden" bias rather than just test for it, assuming that assignment to Control Group 1 and the treatment group is strongly ignorable. In general, if there is evidence under specified assumptions to test for bias, that evidence can instead be used to improve inference. This adjustment also relates to the ideas of reference sampling or substitution sampling, where samples are taken at later points in time to compare to earlier groups and thereby create adjustments (Hirano, Imbens, Ridder, & Rubin, 2001; Rubin & Zanutto, 2002).

1.4. The School Dropout Demonstration Assistance Program

This article was motivated by an applied problem in which the originally chosen control group has insufficient covariate overlap with the treatment group. The School Dropout Demonstration Assistance Program was an initiative operating between 1991 and 1996 in 85 schools, financed by the Department of Education, to help determine effective strategies to reduce school dropout. Here we focus on the "restructuring" programs, which treated entire schools, putting in place structures and services designed to affect all students in the school, such as curriculum reform or expanded teacher training. As one example, the Grand Rapids, Michigan, high school restructuring effort was to adopt a ninth-grade program organized around "family groups," block scheduling, and interdisciplinary themes as well as student services such as student advocates, social workers, and substance abuse specialists (Agodini & Dynarski, 2004; Dynarski, Gleason, Rangarajan, & Wood, 1998). Five restructuring programs were chosen to be part of the evaluation of program impacts; these were located in Dallas, Grand Rapids, Philadelphia, Phoenix, and Santa Ana. A comparison school in the same school district was chosen for each of these schools. We concentrate on the restructuring program in Grand Rapids to illustrate ideas; as such, this will be our treatment group.

We focus on a subset of the covariates that were collected, those deemed by Agodini and Dynarski (2004) to be potentially related to baseline values of four outcomes (dropping out, educational aspiration, absenteeism, and self-esteem). These covariates, X, and their means in the restructured and comparison high schools in Grand Rapids are listed in Table 1. This is a fairly complete list, although as discussed by Agodini and Dynarski, the list is arguably missing some important covariates related to dropping out, such as family income or measures of socioeconomic status other than parents' education level, not available in the SDDAP data.

As seen in Table 1, there are large differences in the means of student-level covariates between the restructured and comparison high schools in Grand Rapids. We show two measures of the imbalance: (a) the absolute standardized bias, defined as the absolute value of the difference in means divided by the standard deviation in the original control group, and (b) the p value of a test of whether the means are different in the treatment and control groups. Absolute standardized biases greater than 0.25 are considered to be particularly problematic (Ho et al., 2007); 3 absolute standardized biases exceed that value, and 10 of the 34 covariate means are significantly different at the 5% level in the treatment and comparison groups. A more thorough investigation of balance should also compare other characteristics of the covariate distributions in the two

TABLE 1.
Summary of Covariates: Two Grand Rapids High Schools

Covariate	Restructured School Mean	Comparison School Mean	Absolute Standardized Bias[a]	*p* Value[b]
Key risk factors (%)				
Does not live with both parents	37	35	0.05	.52
On public assistance	9	6	0.11	.18
Primary language not English	3	2	0.11	.24
Overage for grade	22	22	0.00	1.00
Low course grades in past year	**17**	**23**	**0.15**	**.03**
Discipline problems	40	41	0.01	.91
External locus of control	32	30	0.03	.67
Has own child	3	3	0.00	1.00
Female (%)	51	52	0.01	.90
Race (%)				
Hispanic	**60**	**39**	**0.43**	**.00**
African American	**34**	**54**	**0.40**	**.00**
White	6	7	0.03	.72
Age	15.93	15.93	0.00	.94
Number of schools attended since first grade	4.18	4.20	0.01	.88
Number of siblings	2.55	2.41	0.09	.21
Reading test score (0–100)	**49.42**	**53.84**	**0.27**	**.00**
Math test score (0–100)	**50.09**	**52.28**	**0.16**	**.02**
Aspire to earn a BA (%)	80	76	0.11	.11
Expect to earn a BA (%)	68	63	0.10	.15
Expect to graduate from high school (%)	98	99	0.07	.53
Self-esteem (continuous measure)	**0.32**	**0.20**	**0.16**	**.01**
Talked with parents about school in last year (%)	**95**	**91**	**0.13**	**.04**
Absent more than 10 days in last school year (%)	**29**	**36**	**0.14**	**.04**
Ever skip school (%)	**35**	**46**	**0.23**	**.00**
Ever late for school (%)	60	59	0.01	.91
Active in extracurricular activities (%)	84	79	0.12	.08
School climate (continuous measure)	–0.03	–0.05	0.03	.67

(continued)

Table 1 (continued)

Covariate	Restructured School Mean	Comparison School Mean	Absolute Standardized Bias[a]	p Value[b]
Do > 1 hour of homework per week (%)	72	74	0.05	.51
Read for fun more than 1 hour per week (%)	61	60	0.00	1.00
Watch TV more than 3 hours per day (%)	**53**	**40**	**0.25**	**.00**
Mother has bachelor's degree (%)	25	20	0.11	.10
Father has bachelor's degree (%)	30	26	0.02	.18
Sibling has dropped out of school (%)	15	17	0.06	.46
Student previously dropped out of school (%)	3	3	0.00	1.00
Sample size[c]	428	434		

a. Absolute standardized bias defined as the absolute value of the difference in means divided by the standard deviation in the control group: $\frac{|\bar{X}_t - \bar{X}_{c1}|}{\sigma_{c1}}$. Self-esteem and school climate variables centered to have mean 0 and variance 1 in full population.

b. Variables significantly different at a .05 level are marked in bold. Based on t-statistics for continuous variables and χ^2 test statistics for categorical variables.

c. Means, absolute standardized biases, and p values based on available cases, so exact sample size for each variable may differ.

groups: variances, correlations, and so on (Ho et al., 2007; Rubin, 2001). However, in the SDDAP data even the means are not similar, indicating a problem.

Because there is limited covariate overlap between the two groups, estimation of the unobserved potential outcomes using standard methods would rely heavily on underlying modeling assumptions, due to the extrapolation that would be required. Standard matching methods also would not be useful here because there are an insufficient number of potential matches in the local comparison school. To address this problem, we propose the formation of a comparison "pseudoschool," composed of students from multiple comparison schools. Thus, the primary control group (C1) comprises children in the untreated local comparison school chosen by the SDDAP evaluation. The second control group (C2) comprises students in the other comparison schools with reliable data (those in Dallas, Phoenix, and Santa Ana). By using this second source of comparison students, we can obtain better matches on the X covariates than if we had to obtain matches from C1 for all of the treated students. In particular, we address how to use information from both control sources: students from the local comparison school, who have relatively limited overlap with the treated students on observed student-level covariates, and students who are close matches on these observed covariates but who are from nonlocal comparison schools.

Another possible way to use comparison students from multiple schools would be to simply pool them all together into one large comparison group and estimate the propensity score using the treatment group and pooled comparison group, with an indicator variable for the area in which each student lives included as a covariate. However, this will drive the propensity score specification in an undesirable way, essentially allowing only matches from the local area, especially if all of the treatment group is from one area and relatively few of the comparison students are from that area, as is true in this example. That is, in such cases, the propensity score will essentially equal the indicator variable for local/nonlocal. We would like to obtain exact matches on the area variable when possible but not at the expense of close matches on all of the other covariates; in some sense, we treat the area indicator as a "special" matching variable.

2. Matching With Two Control Groups: Designing the Observational Study

2.1. Obtaining Matches From Both Control Groups

Perhaps the first question that arises when designing a matched observational study is how to choose the matches from the two control groups. Here we discuss extended caliper matching, which is related to the ideas underlying caliper matching (Althauser & Rubin, 1970; Cochran & Rubin, 1973; Rosenbaum & Rubin, 1983). Stuart (2004) proposed another method that fixes the proportion of matches from one control group, but results in that work indicated that extended caliper matching has better performance, and thus we discuss that method here.

In the SDDAP context, for each student in the restructured (treatment) school, if there is a local match within a fixed caliper or "distance" (e.g., within 0.25 standard deviations of the treatment group's propensity scores), the closest local control student is chosen. If there are no local matches within that "Group 1 caliper," then the closest match from outside the local area is taken. Different Group 1 calipers generate different numbers of local versus nonlocal matches. Large Group 1 calipers indicate a preference for local matches: As the Group 1 caliper approaches infinity, a local match is taken regardless of how close (or far apart) the nonlocal matches are from the treatment group. Smaller Group 1 calipers correspond to greater tolerance for nonlocal matches. At the extreme, a Group 1 caliper of 0 indicates that local matches have no priority; the closest match on X is taken, regardless of which control group it is from. Generally, the extended caliper matching procedure is designed to ensure that a match is chosen from outside the local area only when a close local match cannot be found;

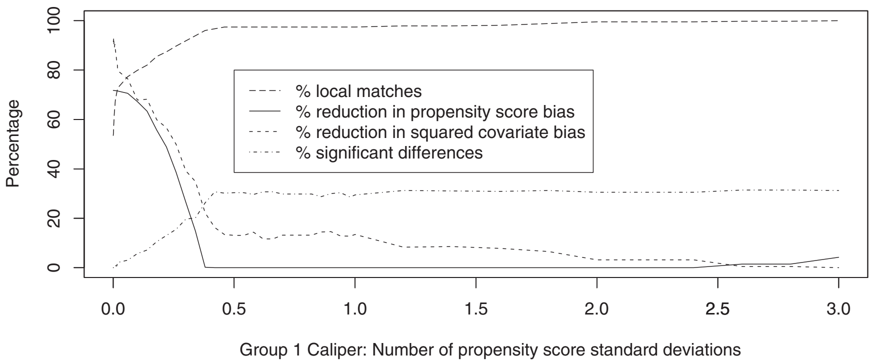

FIGURE 2. *Results from extended caliper matching in School Dropout Demonstration Assistance Program.*

the nonlocal students (Control Group 2) are used only as much as "necessary" while still ensuring close covariate matches on all observed covariates.

The extended caliper matching method was implemented using the SDDAP Grand Rapids High School data using the one-dimensional distance defined by the linear propensity score, estimated using all covariates listed in Table 1, and all units from the treatment group and both control groups. Due to difficulties when including an area indicator in the propensity score model, as discussed in Section 1.4, for the propensity score estimation, the units from both control groups are pooled as if from one large control group. There are theoretical reasons supporting such pooling for this estimation (Rubin & Stuart, 2006). The effects of matching for these data are summarized in Figure 2.

The percentage reduction in bias is defined for quantity B as $100 * \frac{B_m - B_f}{B_f}$, where B_m is the bias in the matched samples and B_f is the bias in the full samples. For the propensity score bias between Groups 1 and 2, $B = \overline{e}_1 - \overline{e}_2$, where e denotes the estimated propensity score. For the squared covariate bias between Groups 1 and 2, we use a multivariate generalization of the standardized bias defined in Table 1: $B = (\overline{X}_1 - \overline{X}_2)' \Sigma^{-1} (\overline{X}_1 - \overline{X}_2)$, where Σ is the variance-covariance matrix of X in the treatment group.

As expected, the maximum bias reduction in the propensity score is obtained with a Group 1 caliper of 0, which takes the closest propensity score match for each treated student, regardless of whether the match is local or nonlocal. This results for the SDDAP data in approximately 55% of the matches being from the local area, indicating that for approximately half of the treated students, their "best match" is in the local comparison school but that there is also the need for some matches from outside the local area to obtain well-matched samples overall. In this example, the bias reduction decreases dramatically for larger Group 1 caliper sizes. More than 90% reduction in squared bias (from 0.49 to 0.04) is obtained when the Group 1 caliper is 0, whereas for a Group 1 caliper of half of a standard deviation or larger, the reduction in squared bias is less than 20%. Group 1 calipers larger than 0.5 of a standard deviation lead to essentially only local matches being chosen, which severely limits the bias reduction on *X* that is possible.

2.2. Choosing the Group 1 Caliper Size

In Section 2.1, the quality of the matches was assessed by taking into account only the observed covariates used to estimate the propensity score. However, a key concern may be that by including matches from outside the local area, we could be introducing bias due to area-level differences: Students in Grand Rapids may be different from students in Dallas or Santa Ana or Phoenix on some unobserved covariate such as community attitudes about dropouts. Assessing the reasonable percentage of matches from each group should thus consider the possible introduction of bias that may result from including matches from outside the local area. For concreteness, we continue to discuss this issue in the context of the SDDAP.

In particular, previous empirical research (Glazerman, Levy, & Myers, 2003; Heckman, Hidehiko, & Todd, 1997; Heckman, Ichimura, & Todd, 1998) indicates that having local area matches is very important for replicating experimental results using observational data, at least in the context of job training programs. Here we provide a way to trade-off that importance with the importance of obtaining close matches on individual-level covariates. The trade-off involves asking questions such as “Would I rather match a student from Grand Rapids to another student from Grand Rapids who is vastly different from the original student in terms of test scores and parents’ education or to a student from Dallas who has identical test scores and parents’ education as the student of interest?” Of course, we do not know the answer to such questions, which are substantive and depend on the applied setting and require the advice of experts, but we attempt to provide a way to make use of that expertise.

We also note that it can be reasonable to assume that there is no additional bias created by obtaining matches from the second control group (e.g., by obtaining matches from outside the local area), even in settings where area differences could be important. For example, Dehejia and Wahba (1999) found that they were able to replicate well the results from a randomized experiment estimating the effect of a job training program in New Jersey using matched observational national data sets (such as the Current Population Survey), which contain individuals from across the United States, and presumably few, if any, from New Jersey in the matched groups. Even though a priori one might expect that being in or out of New Jersey would be important for predicting posttreatment earnings of New Jersey trainees if they were not trained, in this evaluation, obtaining close matches on the observed individual-level covariates (i.e., income in the 2 years prior to the study, race, marital status, years of education, etc.) removed essentially all of the bias in the estimation of the average treatment effect.

Operationally, the most obvious way to implement extended caliper matching would be to determine the optimal Group 1 caliper size given this trade-off between local and nonlocal matches. However, for the theory and approximations given in the following, we determine the optimal number of matches to obtain from Control Group 1 rather than the optimal Group 1 caliper size. Once the optimal number of matches from Control Group 1 is estimated, the Group 1 caliper size can be adjusted accordingly. This is primarily done for simplification of the calculations and approximations. Although researchers such as Cochran and Rubin (1973) have investigated the bias reduction possible with varying caliper sizes (in the setting with one treatment group and one control group), the approximations in that article assume an infinite number of units in the control group, which is not appropriate for our setting with finite samples from the treatment group and Control Group 1.

2.3. Theoretical Setting

The estimand of interest is τ, defined in Section 1.2. We begin by assuming that (unknown to the investigator) there is no effect of the treatment: $Y_i(0) = Y_i(1) = Y_i$ for all individuals i, and thus $Y_i(0) = Y_i(1) = Y_i$. We consider the standard regression setup for individual i with the expected value of the outcome Y_i a linear combination of one individual-level covariate, X_i, which could be a scalar summary of p covariates, such as the propensity score, and an indicator for the area (or district, in the SDDAP setting), D_i, $D_i = 0/1$ for local/nonlocal:

$$E(Y_i|X,D) = \beta X_i + \delta D_i.$$

We consider the case with one treatment group, two control groups, and covariates normally distributed within each group, where μ_t represents the mean of X in the treatment group, σ_t^2 is the variance of X in the treatment group, and N_t is the sample size in the treatment group. Analogous notation holds in Control Groups 1 and 2, indexed by $c1$ and $c2$, respectively. All individuals in the treatment group and Control Group 1 (the SDDAP local control group) have $D_i = 0$, whereas all individuals in Control Group 2 (the SDDAP nonlocal control group) have $D_i = 1$. We assume that Control Group 2 is infinite in size so that exact matches on X can be found from that group for each of the treatment group members. This assumption of an infinite Control Group 2 is used for ease of computation; in fact, as shown in the following, it is not crucial for the results, and the resulting approximations work quite well even with finite samples. Let m be the number of matches chosen from Control Group 1; we are interested in determining the optimal value of m for given β and δ.

The trade-off to consider is that obtaining close matches on X may result in higher bias when estimating τ due to D, and analogously, obtaining close matches on D may result in higher bias due to X. This trade-off is in fact often the case.

Without loss of generality, we assume that $\mu_t > \mu_{c1}$. Then the matching will essentially match the m students with the smallest values of X in the treatment group to the m students in Control Group 1 with the largest values of X. The remainder of the matches (from Control Group 2, matched to the treated students with the $N_t - m$ largest values of X) will match the remaining treated students' covariate values exactly because Control Group 2 is assumed to be infinite in size.

The expected bias in the estimated treatment effect, $\hat{\tau} = \overline{Y}_t - \overline{Y}_{mc}$, where Y_t and Y_{mc} are the observed outcomes in the treatment and matched control group, respectively, is approximately

$$\begin{aligned}
E(\hat{\tau}) &= \beta E(\overline{X}_t - \overline{X}_{mc}) + \delta E(\overline{D}_t - \overline{D}_{mc}) \\
&= \beta(\mu_t - E(\overline{X}_{mc})) + \delta(0 - \frac{N_t - m}{N_t}) \\
&= \beta\mu_t - \beta\left(\frac{m}{N_t}\mu_{c1} + \frac{m}{N_t}\frac{\pi}{4}\sigma_{c1}\log\left(\frac{N_{c1}}{m}\right) + \frac{N_t - m}{N_t}\mu_t + \frac{N_t - m}{N_t}\frac{\pi}{4}\sigma_t\log\left(\frac{N_t}{N_t - m}\right)\right) \\
&\quad + \delta\left(\frac{m - N_t}{N_t}\right).
\end{aligned}$$

This formula uses the approximation for the tail expectation of a univariate standard normal distribution from Rubin (1976), $\Omega(N, n) \approx \frac{\pi}{4}\ln(\frac{N}{n})$. The value of m that minimizes $E(\hat{\tau})$ is the solution to the equation

$$\log\left(\frac{(N_t - m)^{\sigma_t}}{m^{\sigma_{c1}}}\right) = A, \tag{1}$$

where $A = \frac{4}{\pi}(\mu_t - \mu_{c1}) + \sigma_{c1} - \sigma_t + \frac{4}{\pi}\frac{\delta}{\beta} - \sigma_{c1}\log(N_{c1}) + \sigma_t\log(N_t)$. If the variance of X in the treatment group is the same as the variance of X in Control Group 1 ($\sigma_t^2 = \sigma_{c1}^2$), then $m_{opt} = N_t\frac{1}{1 + (\exp(A))^{1/\sigma_t}}$. Further simplification is obtained if $\sigma_t^2 = \sigma_{c1}^2 = 1$, in which case $m_{opt} = N_t\frac{1}{1 + \exp(\frac{4}{\pi}(\mu_t - \mu_{c1}) + \frac{4\delta}{\pi\beta} + ln(\frac{N_t}{N_{c1}}))}$. If the variances of X are not the same in the treatment group and Control Group 1, then a constrained optimization algorithm such as bisection (Lange, 1999) can be used to estimate the optimal m.

Using an estimate of the $\frac{\delta}{\beta}$ ratio, we can use Equation 1 to estimate the optimal number of matches from each of the two control groups. Consistent with designing an observational study without access to the outcome variables, this quantity should not be estimated using the data from the current study but rather should use prior knowledge or results from previous studies. Simulations to assess the performance of this approximation are reported in Stuart (2004); even though the approximation assumes an infinite Control Group 2, results indicate

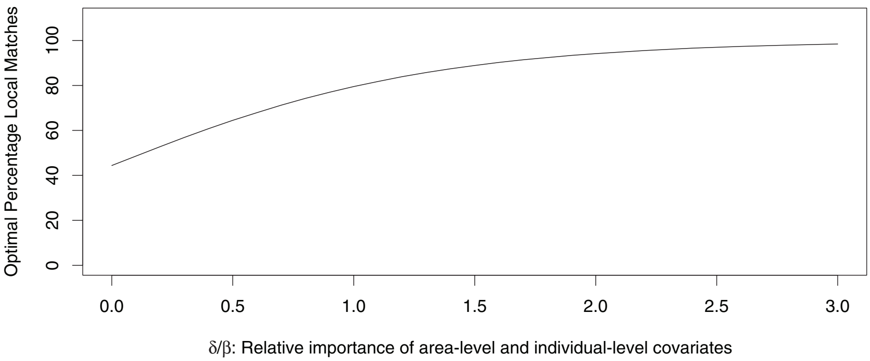

FIGURE 3. *School Dropout Demonstration Assistance Program: Optimal percentage local matches.*

that it holds well even when Control Group 2 is at least twice as large as the treatment group.

2.4. Choice of m in SDDAP

For the SDDAP, we use the results in Section 2.3 to estimate the optimal number of matches from the local control group. Figure 3 shows the optimal percentage local matches for a range of values of $\frac{\delta}{\beta}$, where X is the linear propensity score. If the area-level covariates are not at all important in predicting the outcome ($\delta = 0$), then the optimal percentage local matches is approximately 45%, which is quite close to the percentage matches chosen from the local area with a Group 1 caliper of 0 (55%) from Section 1.2, which essentially assumes $\delta = 0$.

Ideally we would like for this plot to be fairly flat over a range of plausible values of $\frac{\delta}{\beta}$, which would imply that the estimates of the optimal percentage local matches would not be too sensitive to misestimation of this ratio. This result would be especially important when, as in many observational studies, there are many outcome variables not perfectly correlated and there is a desire to use the same matched control group for all outcomes to better replicate the design of a randomized experiment. In the SDDAP example shown in Figure 3, we see that the slope is fairly steep for values of $\frac{\delta}{\beta}$ less than 1.5; however, this will not be true for all data sets.

3. Analysis of the Outcome and Adjusting for Differences Between the Control Groups

One of the benefits of the matching methods described earlier is that they do not make use of the values of the outcome, thus preventing intentional or unintentional bias due to selecting a matched sample to achieve a desired result

(Rubin, 2001). This section describes how to make use of the outcome values of the two control groups (but not those of the treatment group), after the matched samples have been selected, to adjust for differences between the two control groups. In particular, after doing the matching and collecting the outcome data, researchers may want to adjust for a "group effect," capturing potential differences between the control groups that are not reflected by the observed covariates X; here we provide a procedure for doing so. For this theoretical work, as in Section 2, we consider a setting with one observed individual-level covariate X (which again may be a function of p covariates, such as the propensity score) and the indicator D, which distinguishes Control Groups 1 and 2. Using a setup similar to that in Rubin (1973b), let the expected values of the potential outcomes $Y_i(0)$ and $Y_i(1)$ have the following form for individual i with values X_i and D_i:

$$E(Y_i(0)|X,D) = \gamma_c + V(X_i) + (\delta_0 + \delta_1 X_i)D_i, \tag{2}$$

$$E(Y_i(1)|X,D) = \gamma_t + V(X_i) + (\delta_0 + \delta_1 X_i)D_i, \tag{3}$$

where $V(X)$ is an unknown and generally nonlinear but monotone function of X, common to both $Y(0)$ and $Y(1)$. The true average treatment effect is additive, and so $\tau = \gamma_t - \gamma_c$. We refer to the conditional expectations in Equations 2 and 3 as *response surfaces*, using the terminology common in experimental design and used in both Cochran and Rubin (1973) and Rubin (1979), among others.

The intuition behind this method can be seen in Figure 1, which illustrates the scenario for our theoretical situation with one covariate (X). In Figure 1, $\delta_0 = \delta$ and $\delta_1 = 0$ so that there is a constant "district effect" between Control Groups 1 and 2. Because the response surfaces may differ in Control Groups 1 and 2 (as seen in the two distinct parallel response lines in Figure 1), we will adjust the observed outcomes of the matches from Control Group 2 by an estimate of the difference between Control Groups 1 and 2. That difference (δ) is estimated using the group of units from Control Group 1 who look most similar to the treatment group (in the X-space between the two vertical lines) and well-matched units from Control Group 2. The parameter δ could instead be estimated using all units from Control Groups 1 and 2; we instead use only well-matched units from the two groups to prevent reliance on the assumption of linearity of Y given X across the entire covariate space. This adjustment for the "district effect" can be thought of as making the outcomes for the matches from Control Group 2 look as if they "could have been" from Control Group 1.

The outlined procedure uses information from both control groups and attempts to account for the group effect, represented by δ, in the region of the treatment group. The extended caliper matching algorithm described in Section 2 is used to select a set of units from Control Groups 1 and 2 who look the most similar to the treatment group. The potential outcome under control is then imputed for each treated unit. For treated units with a match from Control Group 1, that control unit's outcome value is imputed as the treated unit's outcome value. For treated units with a match from Control Group 2, the match's outcome is used, but the value imputed is adjusted by the estimated difference between Control Groups 1 and 2 (δ). Multiple imputations of the potential outcomes under control are created to account for the uncertainty in estimating δ. Appendix A provides details of the proposed matching adjustment procedure, assuming a normally distributed outcome variable.

The method is expected to work well even when the overall relationship between the covariate X and the outcome ($V(X)$) is linear or nonlinear. Whereas standard ordinary least squares (OLS) adjustment assumes a linear relationship across the entire X distribution in the treatment and control groups, this method assumes linearity only in the area of covariate overlap between Control Groups 1 and 2 (used to estimate δ). However, the basic version of this method does assume that there is no interaction between D and X; that is, D is assumed to have the same additive effect across the entire X distribution. Sensitivity to this assumption is assessed in a set of simulations detailed in Section 4. Results indicate that the matching adjustment method is not particularly sensitive to this assumption.

4. Evaluation of Proposed Adjustment Method

4.1. Simulation Setting

This section summarizes Monte Carlo simulations done to assess the performance of the matching adjustment procedure described in Section 3 relative to no matching and no adjustment. The simulation setting is similar to that in Rubin (1979) and Rubin and Thomas (2000), where matching and OLS are compared in a range of settings with nonlinear response surfaces. A companion article to this one, Rubin and Stuart (2007), extends the simulations in this article to compare the matching adjustment procedure and OLS.

We consider a situation similar to the SDDAP setting, with one treatment group and two control groups. We again assume that Control Group 2 is infinite in size so that exact matches on X can be found from that group. Although the assumption of an infinite Control Group 2 is impossible to satisfy in practice, this setting can still provide guidance for real-world situations: If an infinite second control group does not help much, then it is unlikely that a finite second control group would provide any real assistance, especially considering the additional cost constraints that may make it more expensive to obtain data from a second control group. Section 5 considers the finite Control Group 2 of the SDDAP.

Nonlinear monotone response surfaces that assume a constant treatment effect were examined, with a single covariate X:

$$E(Y_i(j)|X_i, D_j) = \gamma_j + e^{aX_i} + (\delta_0 + \delta_1 X_i)D_j, \tag{4}$$

for group j, $j =$ treatment (t), Control Group 1 (c1), Control Group 2 (c2); $D_t = D_{c1} = 0$; $D_{c2} = 1$. The simulations vary two types of parameters: (a) those that govern the covariate distributions in the groups and thus are essentially known to an investigator and (b) those that are related to the model of the outcome and thus are unknown. The true treatment effect is assumed to be zero, which is no restriction when the treatment effect is additive. We also assume that there is no bias due to any other unobserved covariates; the outcome calculated for each individual is the mean response of each subject conditional on the parameter values, covariates, and control group membership.

Because no residual variance is added to Equation 4 to generate the potential outcomes, we use integrated squared bias (ISB) rather than mean square error as our measure of bias in the estimated treatment effect. The ISB of the estimated treatment effect is defined as $\text{ISB} = (\widehat{\text{ate}} - (\gamma_{\text{t}} - \gamma_{c1}))^2 = (\widehat{\text{ate}})^2$, where $\widehat{\text{ate}}$ is the estimated average treatment effect estimated using the matching adjustment procedure described in Section 3. The initial ISB (IISB) is defined as the squared difference in means of the outcome in the treatment group and Control Group 1 minus the true treatment effect: $\text{IISB} = (\bar{y}_t - \bar{y}_{c1} - (\gamma_t - \gamma_{c1}))^2 = (\bar{y}_t - \bar{y}_{c1})^2$. Because the initial ISB between the groups varies across simulation settings, as in Rubin (1979) we present results in terms of the percentage reduction in ISB, defined as $100 * ((\text{IISB} - \text{ISB})/\text{IISB})$. We summarize here the results of the simulations and present the details in Appendix B.

4.2. Results

4.2.1. Percentage of Matches From Control Group 1

One feature of the extended caliper matching method is that in situations where the researcher does not specify the optimal percentage of the matches from Control Group 1 and instead uses a fixed Group 1 caliper size, the proportion of matches chosen from Control Group 1 will automatically depend on how close the distributions of covariates are in the treatment group and Control Group 1. Using approximations from Rubin and Thomas (1992), for each simulation setting we can calculate the maximum percentage bias reduction possible when matching the treatment group and Control Group 1. Simulation settings with potentially large reductions in bias when matching using just the treatment group and Control Group 1 (e.g., a larger ratio of control units to treated units, or treatment and Control Group 1 distributions with a relatively large amount of overlap) will imply a larger proportion of matches chosen from Control Group 1 rather than Control Group 2.

This relationship is summarized in Figure 4, which shows the percentage of matches chosen from Control Group 1 as a function of the maximum possible bias reduction from matching using just Control Group 1. Settings with maximum possible bias reduction greater than 100% indicate settings where 100% bias reduction is theoretically possible, and thus those are settings that are particularly conducive to matching (Rubin & Thomas, 1996). As expected, when there is a larger potential for bias reduction using just Control Group 1 (particularly

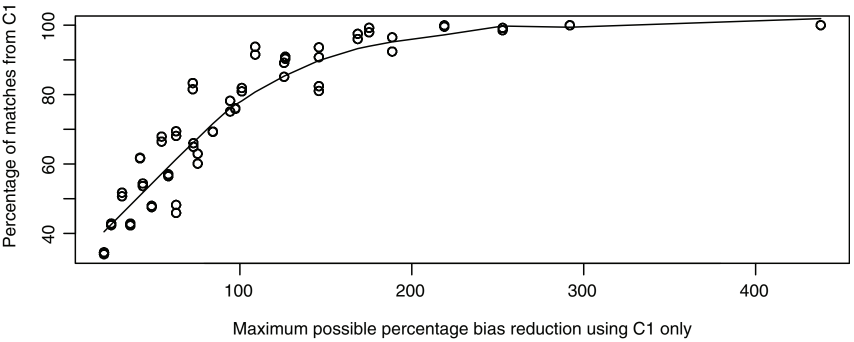

FIGURE 4. *Percentage of matches chosen from Control Group 1 (C1).*

values greater than 100%), more matches are chosen from Control Group 1 rather than Control Group 2. The relationship shown in Figure 4 reflects the fact that when there are more treated units who have a match from Control Group 1 within their caliper, fewer matches are obtained from Control Group 2.

4.2.2. Reductions in Integrated Squared Bias

The matching adjustment procedure yields reductions in ISB for nearly all 1,800 simulation settings examined, with an average percentage reduction in ISB of 80.6%. In particular, larger reductions in ISB are obtained for settings with a smaller value of the coefficient on the interaction between X and D (δ_1), many more units in Control Group 1 than in the treatment group, and larger ratios of the covariate variance in Control Group 1 to that in the treatment group. Details of the percentage reduction in ISB obtained across the simulation settings are given in Appendix B.

Some of these distributional parameters are ones about which a researcher will have some knowledge. In particular, when doing the matching, a researcher will be able to estimate the parameters that describe the covariate distributions. With regard to these parameters, a large percentage reduction in ISB is obtained when the covariate means in the treatment group and Control Group 1 are similar, when the covariate variance in the treatment group is smaller than the covariate variance in Control Group 1, or when there are many more control units than treated units. These results for settings with one treatment group and two control groups correspond to parallel results found in Rubin (1973a) and Rubin and Thomas (1996) for settings with one treatment and one control group.

In contrast, a researcher will not have firm knowledge about the relative sizes of the parameters governing the distribution of Y given X, namely, the response surface (a, δ_1). Larger reductions in bias are achieved when δ_1 is small. This is as expected because the matching adjustment procedure assumes that $\delta_1 = 0$. Thus, some knowledge of whether the difference between Control Groups 1 and 2 varies with the covariate X can help determine whether this adjustment method is suitable. The performance of the procedure depends only somewhat on the value of a, with the method performing the best when $a = -1$. We note that standard OLS estimates would be particularly sensitive to the value of a, performing worse when a is farther from 0. Thus, this matching adjustment procedure appears to be less sensitive than is OLS to nonlinearity in the response function, as further explored in Rubin and Stuart (2007).

5. Adjustment in the SDDAP

5.1. Setup

We will use the SDDAP example to further examine the matching adjustment procedure. We use a simulated outcome variable that is based on a realistic model of an observed outcome, reading score 2 years after the implementation of the restructuring program. The covariate used (X) is baseline reading score. Baseline and outcome reading scores are both on a scale from 0 to 100.

Two response surfaces are considered. These correspond to $V(X)$ in Equations 2 and 3. Parameter values for both were estimated using the observed outcome reading scores, such that both models fit the real data well. The two models are:

1. $E(Y_1|X,D) = a_1 + b_1X + (\delta_0 + \delta_1X)D$
2. $E(Y_2|X,D) = a_2 + e^{b_2X} + (\delta_0 + \delta_1X)D$

where for each value of δ_1 (δ_0 is set to equal 0 throughout), 1,000 random values of the parameters are drawn from the following distributions: $a_1 \sim N(10,5)$, $b_1 \sim N(0.75, 0.125)$, $a_2 \sim N(25, 2.5)$, and $b_2 \sim N(0.0325, 0.005)$. These parameter values resulted in linear R^2 values of 1 for the linear outcome in Model 1 and approximately .85 for the nonlinear outcome in Model 2.

The sample sizes and baseline reading scores are from the SDDAP using Grand Rapids High School as the treated school; only the outcome reading scores are simulated. There are 428 students in the Grand Rapids restructuring school, 434 in the local Grand Rapids comparison school, and 1,111 in the nonlocal comparison schools.

The estimated treatment effect is calculated using the matching adjustment procedure described in Section 3. Again we do not add residual variance to the response surfaces and thus consider the effects of the procedures on ISB. Without loss of generality, we assume that there is no effect of the treatment ($\gamma_t = \gamma_{c1} = \gamma_{c2} = 0$), and thus the outcome models are the same in the treatment and control groups (i.e., $Y_1(0) = Y_1(1)$ and $Y_2(0) = Y_2(1)$). We evaluate the use of the matching adjustment procedure for both of these outcome variables over a range of values of δ_1 from 0 to 0.2. The covariate X is in the scale of 0 to 100, so δ_1X is still a relatively large number. Without loss of generality, for all simulations, $\delta_0 = 0$. Simulation results not reported here verify that when $\delta_1 = 0$, the

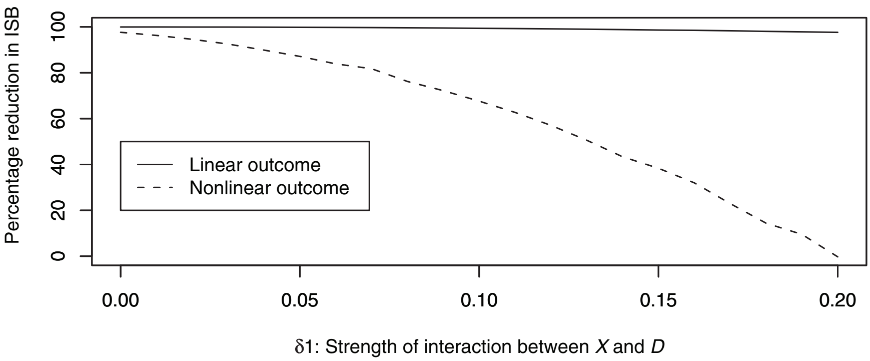

FIGURE 5. *School Dropout Demonstration Assistance Program matching adjustment procedure: Percentage reduction in integrated squared bias (ISB).*

value of δ_0 does not affect the percentage reduction in ISB because δ_0 is well estimated in the group of well-matched controls from both control groups, even in this setting with Control Group 2 of finite size. One hundred sets of simulated outcome values are generated and the full range of δ_1 values are assessed for each data set. With 100 replications, results are accurate to the third decimal place.

5.2. Results

The results from this simulation are summarized in Figure 5. As in Section 4, because of differences in initial bias in the two outcome variables (Y_1 and Y_2), the results are presented as the percentage of initial ISB removed.

When there is no interaction between X and D in the outcome models ($\delta_1 = 0$), the matching adjustment performs very well for both the linear and nonlinear outcomes. With a linear outcome, the percentage reduction in ISB is 99%, and for the nonlinear outcome the percentage bias reduction is 97%. With this data set, using a second control group just three times the size of the treatment group results in more than 95% reduction in ISB, nearly the 100% that would result with an infinite second control group. As the interaction between X and D increases to moderate levels, the performance is still very good for the linear outcome but decreases somewhat for the nonlinear outcome. When the interaction between D and X increases to 0.20, the method does not yield any reduction in bias for the nonlinear outcome. In any applied problem it may be important to think about the potential size of the interaction between the individual covariates (X) and the group indicator (D) when the outcome model is believed to be nonlinear in X.

6. Conclusions

This work has shown the potential for using multiple sources of control units to estimate causal effects. In particular, we have described a method for selecting matched controls from two control groups as well as a procedure to adjust for differences between the groups. The simulations indicate that the method can work very well, even when the assumptions are not fully satisfied. The matching method could be generalized and used for any setting where close matches on some binary covariate are desired, but not at the expense of close matches on the other covariates. Previous matching methods have required a choice between forcing an exact match, simply including the binary variable in the propensity score or Mahalanobis metric, or ignoring the covariate; this work provides a way to consider explicitly the importance of an exact match on that binary covariate.

A companion article (Rubin & Stuart, 2007) extends the analyses reported here, comparing the matching adjustment procedure to standard regression adjustment. Future work should also further examine the optimal percentage of matches to get from each control group and optimal ways of choosing those matches.

Appendix A
Details of Matching Adjustment Procedure

The adjustment method can be implemented using the following procedure, assuming normality of the outcome variable.

1. Match the treatment group and Control Group 1. For this matched group, select only the "good" matches, with "good" defined as being within specified propensity score calipers (Rosenbaum & Rubin, 1985) or a certain percentage of the matches. This group of matched individuals from the treatment group and Control Group 1 is referred to as the C1: T matched group (depicted to the right of the $X = 0$ vertical line in Figure 1).
2. For the individuals in Control Group 1 who are in the C1:T matched group, find matches for them from Control Group 2. Call this the C1: C2 matched group and let $n_{C1:C2}$ be the number of units in this matched group. These units are depicted between the two vertical lines in Figure 1.
3. For the treated units without good matches from Control Group 1 (found in Step 1), that is, the treated units depicted in Figure 1 to the left of the $X = 0$ vertical line, find a match for them from Control Group 2. This is called the C2: T matched group. They are the Control Group 2 units depicted to the left of the left-hand vertical line in Figure 1.
4. Estimate the bias between the two control groups using a model, estimated in the C1:C2 matched group found in Step 2, typically by using a linear model, for example, ordinary least squares (OLS):

$$Y(0)|\beta, \delta, \sigma^2, X \sim N(\beta X + \delta D, \sigma^2 I),$$

   where X consists of the p covariates. Let $\hat{\beta}$, $\hat{\delta}$, and $\hat{\sigma}^2$ be the estimates of β, δ, and σ^2 from this model. In general, βX and δD could be replaced by any nonlinear functions of X and D.

5. In preparation for the imputation of the missing $Y(0)$ values for the treated units, draw (assuming normality)

$$s^2 \sim Inv - \chi^2(n_{C1:C2} - (p+2), \hat{\sigma}^2)$$
$$d|s^2 \sim N(\hat{\delta}, (X^TX)^{-1}s^2).$$

6. For each matched control unit, indexed by k,
If unit k is from Control Group 1 (found in Step 1),

$$\hat{y}_k(0) = y_k(0).$$

If unit k is from Control Group 2 (found in Step 3),

$$\hat{y}_k(0) = y_k(0) - d.$$

In other words, if unit k is from Control Group 2, adjust unit k's outcome by the estimated difference between Control Groups 1 and 2 (d). If unit k is from Control Group 1, leave unit k's observed outcome as is.

7. Create a data set that consists of all treated units' $Y(1)$ values and their matched control units' $Y(0)$ values, with the control outcomes given by $\hat{y}_k(0)$, from Step 6. We then estimate the average effect of the treatment on the treated as $\overline{\boldsymbol{y}(\mathbf{1})} - \overline{\hat{\boldsymbol{y}}(\mathbf{0})}$, where $\boldsymbol{y}(\mathbf{1})$ is the vector of observed values of $Y(1)$ in the treatment group and $\hat{\boldsymbol{y}}(\mathbf{0})$ is the vector of values of $Y(0)$ from Step 6. An extension, explored in a companion article (Rubin & Stuart, 2007) is to use OLS in each imputed data set to obtain a hopefully improved estimate of $\overline{Y(1)} - \overline{Y(0)}$. Here we illustrate the method using the simple difference in means to estimate the treatment effect; however, this step can be modified to run any analysis on the matched data sets (e.g., OLS or a hierarchical model) and the results combined using the multiple imputation combining rules.

8. Repeat Steps 5 through 8 multiple times (i.e., create multiple complete data sets) to represent the uncertainty in the estimation of δ. Use the multiple imputation combining rules (Little & Rubin, 2002; Rubin, 1987, 2004) to obtain an estimate of the average treatment effect and its variance. Specifically, let Q be the average treatment effect, $\hat{Q}_j$ be the estimate of Q found using completed data set j, and U_j be the estimated variance of $\hat{Q} - Q$ found using completed data set j. Generally, let J be the number of imputations (completed data sets) obtained. The multiple imputation estimate of the average treatment effect is $\hat{Q}_{MI} = \frac{1}{J}\sum_{j=1}^{J}\hat{Q}_j$. The estimated variance of $\hat{Q}_{MI} - Q$ is given by $T = \overline{U} + (1 + \frac{1}{J})B$, where $\overline{U} = \frac{1}{J}\sum_{j=1}^{J} U_j$ is the average within-imputation variance and $B = \frac{1}{J-1}\sum_{j=1}^{J}(\hat{Q}_j - \hat{Q}_{MI})^2$ is the between-imputation variance.

Appendix B
Details of Simulations

1. Simulation Setting

Here we provide details of the simulations performed to assess the performance of the matching adjustment procedure. The setting is that of one treatment group and two control groups, with an infinite Control Group 2. We assume that there are N_t units in the treatment group with covariate distribution parameterized such that $X_t \sim N(B/2,\ \sigma_t^2)$ and N_{c1} units in Control Group 1 with covariate distribution parameterized such that $X_{c1} \sim N(-B/2,\ \sigma_{c1}^2)$, where $\frac{\sigma_t^2 + \sigma_{c1}^2}{2} = 1$.

Nonlinear response surfaces were examined, with a single covariate X and a constant treatment effect:

$$E(Y_i(j)|X_i, D_j) = \gamma_j + e^{aX_i} + (\delta_0 + \delta_1 X_i)D_j, \tag{B1}$$

for group j, $j=$ treatment (t), Control Group 1 (c1), Control Group 2 (c2); $D_t = D_{c1} = 0$; $D_{c2} = 1$. The true treatment effect is zero, which is no restriction when the treatment effect is additive. We also assume that there is no bias due to any other unobserved covariates; the outcome calculated for each individual is the mean response of each subject conditional on the parameter values, covariates, and control group membership. The bias in the proposed adjustment method can be seen most clearly by examining mean responses only.

The simulation parameters are the following:

1. Difference between Control Groups 1 and 2: $\delta_0 = 0$; $\delta_1 = 0, 0.25, 0.5, 0.75, 1, 1.25$
2. Treatment group sample size: $N_t = 100, 200$
3. Ratio of Control Group 1 size to treatment group size: $N_{c1}/N_t = 2, 5$
4. Initial bias in X between the treatment group and Control Group 1: $B = 0.5$, 0.75, 1, 1.25, 1.5
5. Variance of X in the treatment group: $\sigma_t^2 = 0.5$, 1, 1.5
6. Amount of nonlinearity in relationship between the response and X: $a = -1$, -0.5, 0, 0.5, 1

The values of the parameters involving the covariate distributions are essentially known to an investigator at the time of matching (e.g., N_t, N_{c1}, B, σ_t^2). The unknown parameters are those involving the true response surface (δ_0, δ_1, a) because generally Y is not yet measured. Without loss of generality, we use only one value for δ_0. In addition, the matching performance is the same regardless of the sign of δ_1, and so only settings with positive values of δ_1 are evaluated. The chosen values of a reflect moderate (± 0.5) and relatively large (± 1) nonlinearity in the relationship between X and Y, as used in Rubin (1973b, 1979). For the range of X distributions considered, a value of a of ± 0.5 generally leads to a linear r^2 value of approximately .85, whereas $a = \pm 1$ leads to a linear r^2 value of approximately .55.

The estimate of the treatment effect using the matching adjustment procedure was obtained as described in the algorithm given in Appendix A. Within each replication, 50 imputed data sets were created. The Group 1 caliper used in the matching adjustment procedure was 0.2 standard deviations of X in the

TABLE B1
Main Effects, Percentage Reduction in Integrated Squared Bias (ISB)

	Percentage Reduction in ISB (Remaining ISB)[a]					
	Factor Levels					
Factor	1	2	3	4	5	6
δ_1	98 (0.01)	98 (0.02)	92 (0.06)	82 (0.12)	67 (0.22)	47 (0.33)
N_t	81 (0.13)	81 (0.12)	NA	NA	NA	NA
N_{c1}/N_t	74 (0.17)	88 (0.09)	NA	NA	NA	NA
B	75 (0.03)	81 (0.07)	83 (0.11)	83 (0.16)	81 (0.27)	NA
a	90 (0.08)	56 (0.10)	90 (0.12)	73 (0.14)	96 (0.19)	NA
σ_t^2	98 (0.02)	91 (0.09)	54 (0.27)	NA	NA	NA

a. The factor values are $\delta_1 = 0, 0.25, 0.5, 0.75, 1, 1.25$; $N_t = 100, 200$; $N_{c1}/N_t = 2, 5$; $B = 0.5, 0.75, 1, 1.25, 1.5$; $a = -1, -0.5, 0, 0.5, 1$; $\sigma_t^2 = 0.5, 1, 1.5$ ($\sigma_t^2 + \sigma_{c1}^2 = 2$).

TABLE B2
Varying δ_1 *and* σ_t^2*, Percentage Reduction in Integrated Squared Bias (ISB; ISB Remaining)*

		δ_1					
		0.00	0.25	0.50	0.75	1.00	1.25
σ_t^2	0.5	99 (0.00)	99 (0.01)	99 (0.01)	98 (0.02)	96 (0.03)	95 (0.05)
	1.0	99 (0.01)	98 (0.02)	95 (0.05)	91 (0.09)	84 (0.15)	76 (0.24)
	1.5	95 (0.02)	95 (0.04)	83 (0.12)	58 (0.26)	20 (0.46)	−30 (0.72)

TABLE B3
Varying B *and* σ_t^2*, Percentage Reduction in Integrated Squared Bias (ISB; ISB Remaining)*

		B				
		0.5	0.75	1.0	1.25	1.5
σ_t^2	0.5	99 (0.00)	98 (0.00)	98 (0.01)	98 (0.02)	96 (0.06)
	1.0	94 (0.02)	93 (0.03)	91 (0.07)	89 (0.12)	86 (0.22)
	1.5	30 (0.09)	52 (0.16)	61 (0.24)	62 (0.35)	63 (0.52)

treatment group. A Group 1 caliper of 0.5 standard deviations was also investigated and bias reduction similar to that presented in Tables B1 to B4 was obtained. With a set Group 1 caliper, the actual percentage of matches from each group will vary depending on how far apart are the distributions of covariates in the treatment group and Control Group 1, as discussed in Section 4.2.1.

TABLE B4

Varying a *and* σ_t^2*, Percentage Reduction in Integrated Squared Bias (ISB; ISB Remaining)*

		a				
		−1	−0.5	0	0.5	1
σ_t^2	0.5	99 (0.02)	98 (0.02)	99 (0.02)	95 (0.01)	96 (0.03)
	1.0	99 (0.06)	86 (0.07)	94 (0.08)	78 (0.10)	95 (0.15)
	1.5	70 (0.15)	−17 (0.22)	76 (0.27)	44 (0.32)	95 (0.40)

Following Rubin (1979) and Rubin and Thomas (2000), for this 1–1 matching we do not add residual variance onto the response surface and thus consider integrated squared bias (ISB) rather than mean square error (the model r^2 is assumed to be 1, and thus the residual variance is assumed to be 0). The ISB of the estimated treatment effect is defined as $\text{ISB} = (\widehat{\text{ate}} - (\gamma_{\text{t}} - \gamma_{\text{c1}}))^2 = (\widehat{\text{ate}})^2$, where $\widehat{\text{ate}}$ is the estimated average treatment effect. The initial ISB is defined as the squared difference in means of the outcome in the treatment group and Control Group 1 minus the true treatment effect:[1] $\text{IISB} = (\bar{y}_t - \bar{y}_{c1} - (\gamma_t - \gamma_{c1}))^2 = (\bar{y}_t - \bar{y}_{c1})^2$.

Generally, 500 replications were run at each setting to compute the integrated squared bias and percentage reduction in ISB of the estimated treatment effect. As in Rubin (1979), the simulation settings are nested when possible (e.g., smaller sample sizes are nested within larger sample sizes) to improve statistical and computational efficiency by inducing correlation between the settings and reducing the number of random numbers drawn. More details and motivation for this type of design are in the appendix of Rubin (1979).

2. Results

An analysis of variance (ANOVA) on the percentage reduction in ISB indicates that δ_1, the ratio of N_{c1} to N_t, a, σ_t^2, and selected interactions all contribute to the variation in percentage reduction in ISB (ANOVA details not presented here). Table B1 summarizes the bias reduction due to the main effects of each factor. Tables B2 to B4 compare the results for selected interactions of the simulation parameters: those indicated by the ANOVA to lead to large variation in the percentage reduction in integrated squared bias.

Note

[1] When the true response surface is linear ($a = 0$) the outcome is constant for all individuals and the difference in means of the outcome in the treatment and control groups is 0. Thus, in that case, the initial integrated squared bias is

defined as the squared difference in covariate means between the treatment group and Control Group 1 because the difference in covariate means approaches the bias in the outcome as the response surface approaches linearity.

References

Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *Review of Economics and Statistics*, *86*, 180–194.

Althauser, R., & Rubin, D. (1970). The computerized construction of a matched sample. *American Journal of Sociology*, *76*, 325–346.

Baker, S. G., & Lindeman, K. S. (1994). The paired availability design: A proposal for evaluating epidural analgesia during labor. *Statistics in Medicine*, *13*, 2269–2278.

Baker, S. G., & Lindeman, K. (2001). Rethinking historical controls. *Biostatistics*, *2*, 383–396.

Campbell, D. (1969). Artifact and control. In R. Rosenthal & R. Rosnow (Eds.), *Artifact in behavioral research* (pp. 351–382). San Diego, CA: Academic Press.

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. *American Economic Review*, *84*, 772–784.

Christakis, N. A., & Iwashyna, T. I. (2003). The health impact of health care on families: A matched cohort study of hospice use by decedents and mortality outcomes in surviving, widowed spouses. *Social Science & Medicine*, *57*, 465–475.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya: The Indian Journal of Statistics, Series A*, *35*, 417–446.

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

Dempster, A. P., Selwyn, M. R., & Weeks, B. J. (1983). Combining historical and randomized controls for assessing trends in proportions. *Journal of the American Statistical Association*, *78*, 221–227.

Dynarski, M., Gleason, P., Rangarajan, A., & Wood, R. (1998). *Impacts of school restructuring initiatives: Final report* (Research report from the School Dropout Demonstration Assistance Program evaluation, Mathematica Policy Research, Inc.). Washington, DC: U.S. Department of Education.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *Annals of the American Academy of Political and Social Science*, *589*, 63–93.

Heckman, J. J., Hidehiko, H., & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, *64*, 605–654.

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies*, *65*, 261–294.

Hirano, K., Imbens, G., Ridder, G., & Rubin, D. B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica*, *69*, 1645–1659.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199–236.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*, 4–29.

Lange, K. (1999). *Numerical analysis for statisticians. Statistics and computing series.* New York/Berlin: Springer-Verlag.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley Interscience.

Lu, B., & Rosenbaum, P. R. (2004). Optimal pair matching with two control groups. *Journal of Computational and Graphical Statistics*, *13*, 422–434.

Morgan, S. L., & Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods & Research*, *35*, 3–60.

Rosenbaum, P. R. (1987). The role of a second control group in an observational study. *Statistical Science*, *2*, 292–316.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). Berlin/New York: Springer-Verlag.

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

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, *39*, 33–38.

Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics*, *29*, 159–184.

Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, *29*, 185–203.

Rubin, D. B. (1976). Multivariate matching methods that are equal percent bias reducing, II: Maximums on bias reduction. *Biometrics*, *32*, 121–132.

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

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, *74*, 318–328.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley.

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

Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety*, *13*, 855-857.

Rubin, D. B. (2006). *Matched sampling for causal inference*. Cambridge, UK: Cambridge University Press.

Rubin, D. B., Cook, S. R., & Stuart, E. A. (2003). *Statistical analysis plan: Assessing the efficacy of Fabrazyme in a Phase 4 study*. Unpublished manuscript prepared for Genzyme Corporation, for submission to the Food and Drug Administration.

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

Rubin, D. B., & Stuart, E. A. (2007). *Using multiple control groups: A comparison of matching adjustment and regression*. Unpublished manuscript.

Rubin, D. B., & Thomas, N. (1992). 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.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, *95*, 573–585.

Rubin, D. B., & Zanutto, E. (2002). Using matched substitutes to adjust for nonignorable nonresponse through multiple imputations. In R. Groves, D. Dillman, R. Little, & J. Eltinge (Eds.), *Survey nonresponse* (pp. 389–402). New York: John Wiley.

Shen, Y., & Fleming, T. R. (1999). Assessing effects on long-term survival after early termination of randomized trials. *Lifetime Data Analysis*, *5*, 55–66.

Smith, H. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology*, *27*, 325–353.

Smith, J., & Todd, P. (2005). Does matching overcome LaLonde’s critique of nonexperimental estimators? *Journal of Econometrics*, *125*, 305–353.

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

Authors

ELIZABETH A. STUART is Assistant Professor, Department of Mental Health and Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; estuart@jhsph.edu; www.biostat.jhsph.edu/~estuart. Her areas of specialty include experimental and nonexperimental designs for estimating causal effects, with particular application to education and mental health.

DONALD B. RUBIN is the John L. Loeb Professor of Statistics, Harvard University, Cambridge, MA 02138; rubin@stat.harvard.edu. His areas of specialty include causal inference, missing data, and Bayesian methods.

Manuscript received January 29, 2006
Revision received December 1, 2006
Accepted April 3, 2007