Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments

William R. SHADISH, M. H. CLARK, and Peter M. STEINER

A key justification for using nonrandomized experiments is that, with proper adjustment, their results can well approximate results from randomized experiments. This hypothesis has not been consistently supported by empirical studies; however, previous methods used to study this hypothesis have confounded assignment method with other study features. To avoid these confounding factors, this study randomly assigned participants to be in a randomized experiment or a nonrandomized experiment. In the randomized experiment, participants were randomly assigned to mathematics or vocabulary training; in the nonrandomized experiment, participants chose their training. The study held all other features of the experiment constant; it carefully measured pretest variables that might predict the condition that participants chose, and all participants were measured on vocabulary and mathematics outcomes. Ordinary linear regression reduced bias in the nonrandomized experiment by 84–94% using covariate-adjusted randomized results as the benchmark. Propensity score stratification, weighting, and covariance adjustment reduced bias by about 58–96%, depending on the outcome measure and adjustment method. Propensity score adjustment performed poorly when the scores were constructed from predictors of convenience (sex, age, marital status, and ethnicity) rather than from a broader set of predictors that might include these.

KEY WORDS: Nonrandomized experiment; Propensity score; Randomized experiment; Selection bias

1. INTRODUCTION

Randomized experiments can yield unbiased estimates of effect sizes. But randomized experiments are not always feasible, and other times ethical constraints preclude random assignment. Consequently, researchers often use nonrandomized experiments (Rosenbaum 2002; Shadish, Cook, and Campbell 2002) in which participants self-select into treatments or are selected nonrandomly to receive treatment by an administrator or service provider. Unfortunately, whatever feasibility or ethical benefits sometimes accrue to nonrandomized experiments, they yield effect estimates that either are demonstrably different from those from randomized experiments (Glazerman, Levy, and Myers 2003) or are at best of unknown accuracy (Rosenbaum 2002). To explore the accuracy of estimates from nonrandomized experiments, previous research has compared randomized and nonrandomized experiments in one of three ways: computer simulations, single-study comparisons, or meta-analysis. All three approaches have weaknesses that the present study remedies. A fourth method that we discuss, the doubly randomized preference trial, works well in theory but in practice is plagued by problems of attrition and partial treatment implementation.

Computer simulations (e.g., Drake 1993) investigate these issues by generating precisely controlled but artificial data, varying key features that might affect results, such as the magnitude of the bias or the sample size. The high control and the large number of replications in these simulations yield very accurate results. But such simulations are quite artificial, for example, presuming that data are normally distributed or that outcome measures have no measurement error. Most importantly, simulations require the researcher to specify the selection model for nonrandomized experiments; but in nonrandomized experiments, the problem is that the researcher does not know that model. So simulations can only approximate real-world selection bias problems, and they do so to an uncertain degree.

Two other methods provide more realistic contexts for studying selection bias (Shadish 2000). The single-study approach compares results from an existing randomized experiment with results obtained when a single nonrandomized control that is conveniently available is substituted for the original randomized control (or, alternatively, by comparing the randomized control with the nonrandomized control on the assumption that if the two control groups are equal, then the nonrandomized control can be substituted for the randomized control). This method gives the researcher access to raw data from individual participants, so that he or she can apply statistical adjustments to those data to improve the estimates. The results of such studies have been mixed, with some studies supporting the use of adjustments and others not doing so; for example, Heckman, Ichimura, and Todd (1997) randomly assigned applicants to a control group or to a job training program, and also collected data on a group of eligible nonparticipants who met the requirements for the training program but were not participating in it. They then compared the randomized treatment group both to the nonrandomized control group (the nonrandomized experiment) and to the randomized control group (the randomized experiment). The two experiments yielded different estimates when adjusted using econometric selection bias models. In comparison, more optimistic results were obtained in studies by Dehejia and Wahba (1999) and Hill, Reiter, and Zanutto (2004) using propensity score adjustments. Hill et al. (2004) also used multiple imputation to cope with the inevitable missing data that occur both before and after treatment in field experiments.

At first glance, studies like those of Dehejia and Wahba (1999), Heckman et al. (1997), and Hill et al. (2004) seem to provide a credible test of the effects of adjustments such as

William R. Shadish is Professor, Psychological Sciences Section, School of Social Sciences, Humanities, and Arts, University of California, Merced, CA 95344 (E-mail: *wshadish@ucmerced.edu*). M. H. Clark is Assistant Professor of Psychology, Department of Psychology, Southern Illinois University, Carbondale, IL 62901 (E-mail: *mhclark@siu.edu*). Peter M. Steiner is Assistant Professor, Institute for Advanced Studies, 1060 Vienna, Austria, and currently a Visiting Research Scholar at Northwestern University, Evanston, IL 60208 (E-mail: *steiner@ihs.ac.at*). Shadish and Steiner were supported in part by grant 0620-520-W315 from the Institute for Educational Sciences, U.S. Department of Education.

Journal of the American Statistical Association
December 2008, Vol. 103, No. 484, Applications and Case Studies
DOI 10.1198/016214508000000733

propensity score analysis or selection bias modeling. However, these studies all share a key weakness that renders their results unclear—they confound assignment method with other study features. These confounds are problematic. Adjustments such as propensity score analysis are attempting to estimate what the effect would have been had the participants in a nonrandomized experiment instead been randomly assigned to the same conditions using the same measures at the same time and place. The latter counterfactual cannot be observed directly. As has been argued in causal inference in general (Rubin 1974; Holland 1986), the best approximation to this true counterfactual may be a group of participants whose assignment method (random or nonrandom) was itself randomly assigned to them, with all other features of the experiment held equal. This was not done by Dehejia and Wahba (1999), Heckman et al. (1997), or Hill et al. (2004), or in any other such studies. Rather, assignment mechanism (random or nonrandom) was varied nonrandomly in those studies and always was confounded with other differences between the random and nonrandom control groups. For example, compared with the randomized control group, the nonrandomized control group often was assessed at different sites or times, by different researchers, with different versions of the measure; and the groups may have had different rates of treatment crossover and missing outcome data. Even if these confounding factors were known, it would be impossible to adjust for some of them, because the single-study approach relies on just one instance of a randomized control and a nonrandomized control, so there is no variability in study-level confounding factors. Consequently, if research that uses the single-study approach finds that a selection bias adjustment to the nonrandomized experiment does (or does not) yield the same results as the randomized experiment, then we cannot know whether this is due to the adjustment method or to variability caused by these other confounding factors.

Meta-analysis offers a partial remedy to the problem of confounding factors by comparing many randomized and nonrandomized experiments on the same question to see whether they yield the same average effect size. Lipsey and Wilson (1993) used the simplest form of this approach, summarizing results from dozens of meta-analyses comparing randomized and nonrandomized experiments. The average over these comparisons was 0—nonrandomized experiments yielded the same effect size as randomized experiments on average—although in any given meta-analysis, the difference usually was not 0. But the validity of this overall average relies on the assumption that any variables that are confounded with assignment method are distributed randomly over meta-analyses. Data suggest that this is unlikely to be the case (e.g., Heinsman and Shadish 1996). In an attempt to lessen reliance on this assumption, other meta-analyses have coded such confounding factors and included them as covariates to get an adjusted difference between randomized and nonrandomized experiments (e.g., Heinsman and Shadish 1996; Shadish and Ragsdale 1996; Glazerman et al. 2003). These meta-analyses have yielded mixed results, with some concluding that the adjusted difference is near 0 (Heinsman and Shadish 1996) and others concluding that it is not (Glazerman et al. 2003).

Fundamentally, however, the meta-analytic approach suffers from the same flaw as the single-study approach, which is not surprising because it is based on those single studies. Variables confounded with assignment mechanism are still unknown, and so the researcher cannot be sure that all relevant confounding covariates have been identified, measured well, and modeled properly. Moreover, the meta-analytic approach also cannot access primary raw data from each experiment, so it cannot test whether such adjustments as selection bias modeling or propensity score analysis improve estimates from nonrandomized experiments.

To address some of the problems with these methods, the present study explores the differences between randomized and nonrandomized experiments using a laboratory analog that randomly assigns participants to be in either randomized or nonrandomized experiments that otherwise are equal in all respects. This equating of experimental methods on conditions other than assignment method remedies the key weakness of both the single-study approach and the meta-analytic approach in which other variables can be systematically confounded with estimates of the effects of assignment method. The method also remedies the additional problem of the meta-analytic approach by producing data on individual participants, allowing the use of adjustments to reduce bias that are not available to the meta-analytic approach. Finally, the method examines naturally occurring selection biases in which the selection process is unknown, a more realistic test than in computer simulations.

The approach in the present study is related to a fourth method—the doubly randomized preference trial (DRPT) (Rücker 1989; Wennberg, Barry, Fowler, and Mulley 1993; Janevic et al. 2003; Long, Little, and Lin 2008)—although it differs in some important ways. First, some of the DRPT literature makes only hypothetical proposals about the possibility of implementing DRPTs (e.g., Wennberg et al. 1993) or is devoted only to developing a statistical model for assessing effects in DRPTs rather than to gathering experimental data with a DRPT (e.g., Rücker 1989). This is nontrivial, because the practical problems involved in executing DRPTs are formidable and, as we argue later, usually impede the ability of DRPTs to obtain a good test of the effects of adjustments, such as propensity score analysis. Second, none of the DRPT studies conducted to date has used the design to assess whether adjustments to observational studies like propensity score analysis can replicate results that would have been obtained had participants been randomized.

Third, and perhaps most importantly, because our method uses a brief laboratory analog treatment, it avoids problems of partial treatment implementation and of missing outcome data that occurred in the few past DRPTs that actually tried to gather data. This is crucial, because adjustments like propensity score analysis answer questions only about what would have happened to the participants in the nonrandomized experiment had they been randomly assigned to conditions. They do not adjust for partially implemented treatments or for missing outcome data, but any DRPT conducted in a field setting is almost certain to encounter both of these problems. For example, nearly two-thirds of those initially assigned to the conditions of Janevic et al. (2003) refused to accept their random assignment to the randomized or choice arms of the study, and all of them had missing outcome data. Although the differential rate of refusal (3%) to conditions was minimal (62% refusal to the choice arm vs. 65% to the randomization arm), an additional 4% withdrew

from the choice arm after the pretest, making the differential missing outcome data $65\% - 58\% = 7\%$. Moreover, such biases might be differential in substantive nature across conditions if those willing to accept no choice of condition (i.e., random assignment) are different from those who are willing to participate only if they can choose their conditions.

Janevic et al. (2003) also reported large and significant differences in treatment implementation rates between the randomization and choice arms of the study. The reanalysis of these data by Long et al. (2008) used an intent-to-treat analysis to estimate causal effects in the presence of such problems, but that analysis cannot be done without additional assumptions beyond an adjustment for assignment method. Thus the resulting comparison of the adjusted results from the randomized and nonrandomized experiments of Janevic et al. (2003) is a joint test of the effects of adjusting for assignment method, missing outcomes, and partial treatment implementation. Our method substantially avoids these two problems and thus allows testing of the effects of adjustments for assignment method that are less encumbered by extraneous concerns.

Our method has its own problems, however. What may be gained in purity of the adjustment for assignment method using the present method may be lost in questions about generalization from the laboratory to the field, about the substantive importance of the brief intervention, and about other issues that we describe in more detail in Section 4. In addition, our method represents only one kind of observational study, a prospective nonrandomized experiment in which participants agree to be recruited and to be randomized to randomization or choice conditions. Those who agree to be recruited to such an experiment may differ from those who self-select into a program of their own accord, as might be more common in retrospective observational studies. Thus the present method is just one alternative with its own strengths and weaknesses compared with past methods.

Nonetheless, the unique contribution of the present study is the novel methodology for testing the accuracy of proposed statistical solutions to a critically important problem in statistical practice. Although at first glance there may be little motivation for interest in a brief laboratory analog treatment, this format is a key virtue, because it allows estimation of the effects of adjustments for nonrandom assignment unconfounded with assumptions about missing outcome data, partial treatment implementation, or other differences between the randomized and nonrandomized experiments. Although one might imagine a field experiment with similar virtues, such as a very brief medical intervention that is fully implemented with an outcome that is a matter of public record and in which participants readily agree to be randomly assigned to whether or not they get a choice of treatment, such a field experiment has yet to occur, and its practical logistics would be formidable.

The rest of this article is organized as follows. Section 2 describes the method and its implementation. Section 3 presents the results, with particular focus on propensity score adjustments. Section 4 discusses the promise and the limitations of this study and suggests ways of extending this methodology to explore its generalizability.

2. METHODS

The study began with baseline tests that were later used to predict treatment selection (Fig. 1). Then participants were randomly assigned to be in a randomized experiment or a nonrandomized experiment. Those assigned to the randomized experiment were randomly assigned to mathematics or vocabulary training. Those who were assigned to the nonrandomized experiment chose which training they wanted and then attended the same training sessions as those who were randomly assigned. After training, all participants were assessed on both mathematics and vocabulary outcomes. This design ensured that all participants were treated identically in all respects except assignment method.

2.1 Participants

Volunteer undergraduate students from introductory psychology classes at a large mid-southern public university were assigned randomly to be in a randomized ($n = 235$) or a nonrandomized ($n = 210$) experiment, using month of birth for practical reasons. These sample sizes were not large, a limitation if propensity scores are most effective with large samples. But such sample sizes are common in applications of propensity scores in field experimentation. Students received experimental credit that was either required or allowed for their classes, and they chose to participate in this experiment from among several available experiments. Of the 450 students who signed up for the experiment, 445 completed the pretests, intervention, and posttests. The remaining five participants dropped out after being assigned to conditions but during the transition from pretest administration to training. Of these, three were randomly assigned to the randomized experiment (two then randomly assigned to mathematics and one to vocabulary), and two were randomly assigned to the nonrandomized experiment (one chose vocabulary, and one did not complete the choice form). These five students were dropped from analyses because their missing outcomes were only 1.1% of the data and because their distribution was even over assignment to random versus nonrandom experiments. These five were the only participants lost to treatment or outcome measurement.

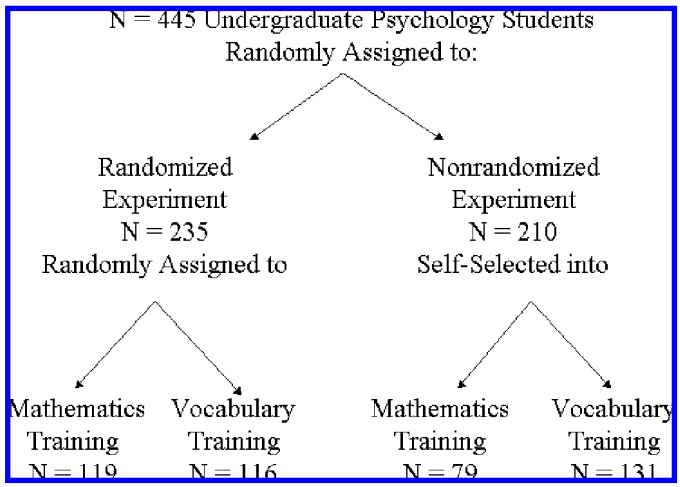

Figure 1. Overall design of the study.

2.2 Pretests

Written instructions and computer scored answer sheets were used for all of the following pretests:

- Demographics Questionnaire I, prepared by us, gathered data about participant age, education, marital status, major area of study, ACT and SAT scores, and grade point average (GPA) for college and high school.
- The Vocabulary Test II (Educational Testing Services 1962) measured vocabulary skills to predict selection into mathematics or vocabulary training.
- The Arithmetic Aptitude Test (Educational Testing Services 1993), administered with scratch paper, measured mathematics skills to predict selection into conditions.
- Demographics Questionnaire II, prepared by us based on an interview with a full-time staff member of the student educational advising center, assessed previous scholastic experiences in mathematics and vocabulary to predict selection into condition.
- The International Personality Item Pool test (Goldberg 1997) assessed five major domains of personality: extroversion, emotional stability, agreeableness, openness to experience, and conscientiousness.
- The Short Mathematics Anxiety Rating Scale (Faust, Ashcraft, and Fleck 1996) assessed stress induced by mathematics to predict selection into mathematics training.
- The Short Beck Depression Inventory (Beck and Beck 1972) assessed depression, given that a previous scale assessing depression in college students (Kleinmuntz 1960) predicted performance.

2.3 Treatments

A series of overhead transparencies presented interventions to teach either 50 advanced vocabulary terms or 5 algebraic concepts. The vocabulary transparencies each included a novel term, its phonetic spelling, and a sentence in which the word was used. The mathematics transparencies included five rules for transforming exponential equations and several examples in which those rules were applied to algebraic formulas. We compared two treatment conditions (rather than comparing treatment to no treatment) for two reasons: (a) Doing so created two effect estimates, one for the effects of vocabulary training on vocabulary outcome and one for the effects of mathematics training on mathematics outcome, and (b) a "no treatment" control might attract a disproportionate number of participants to select the least time-consuming session in the nonrandomized experiment. We chose to train participants in mathematics and vocabulary for three reasons. First, various kinds of mathematics and language skills are studied from elementary school through college, are often used in educational testing and are basic skills for many academic and career fields, so they are good analogs to topics sometimes studied in field experiments. Second, through experimental control over the difficulty of the vocabulary terms and algebraic concepts, we could anticipate that most participants would not be familiar with the material before the experiment and, correspondingly, anticipate that the experimental effect size would be meaningfully large. Third, college students differ greatly in their propensity to choose mathematics training, reflecting a condition ripe for selection bias, thus making it easier to detect differences between randomized and self-selected conditions.

Training sessions were conducted by one of four white males, including three psychology graduate students and one undergraduate psychology major. Trainers were counterbalanced for each trial session and type of training, so that trainers varied what they taught from session to session. Each trainer conducted five or six training sessions in either vocabulary or mathematics. To further standardize testing and treatment conditions across sessions, all training and other instructions were read from a well-rehearsed script.

2.4 Posttest

A 50-item posttest contained 30 vocabulary items (15 items presented in training and 15 new items) and 20 mathematics items (10 presented earlier and 10 new), presenting vocabulary first and mathematics second for all participants in all conditions. This posttest was given to all participants regardless of training. We later found that the correct response for two mathematics items was not listed, however, so those items were removed from analyses.

2.5 Procedure

Data collection spanned 22 weeks, with 24 testing sessions having between 7 and 48 people per session. Participants signed up for the experiment between 4 weeks to 1 hour before participating. On arrival, participants completed consent forms and the Demographics Questionnaire I. The consent form included the option to allow researchers to access university records of their high school GPAs, college GPAs, mathematics and English grades, and ACT or SAT college admission scores; 92% of the participants consented. But university records reported ACT scores for only 61.5% of the participants. Having missing data on this variable was not significantly related to the condition to which the participant was later assigned ($\chi^2 = 1.614$, $p = .204$). We substituted self-reported SAT, ACT, and GPAs for those participants who did not consent or who had missing data in university records, and we converted SAT scores to ACT estimated scores using tables provided by ACT and Educational Testing Services (Dorans, Lyu, Pommerich, and Houston 1997). Although it is possible to estimate missing ACT scores using imputation (e.g., Hill et al. 2004), using self-reported ACT scores is transparent and seemed adequate for present purposes. The remaining pretest materials were then distributed.

Although virtually no outcome data were missing, some data on pretreatment covariates were missing for some participants: 132 (62%) of the quasi-experimental participants had complete predictor data, 37 (18%) had missing data on 1 predictor, and 41 (19%) had missing data on more than 1 predictor. But the overall number of missing observations was quite low (3.11% and 4.2% of all covariate measurements of the randomized and quasi-experiments). Therefore, to maintain the focus on the simple comparison of randomized and nonrandomized evaluations, we filled in missing values using EM-based imputation using the missing-data module of SPSS 14.0. These imputations are biased because they do not include an error component. In subsequent research, we intend to examine the sensitivity of propensity score analyses to different ways of treating missing data.

At the end of the time allotted for pretests, participants were assigned randomly to be in a randomized ($n = 235$) or a nonrandomized ($n = 210$) experiment using randomly chosen months of birth; these randomly chosen birth month assignments were counterbalanced over each training session. Participants born in three randomly chosen months were sent to the vocabulary training condition of the randomized experiment ($n = 116$). Participants born in three other randomly chosen months were sent to the mathematics training condition of the randomized experiment ($n = 119$). As they left for the training sessions, these participants were given packets labeled "R" (for randomized experiment) containing posttest materials. Next, the 210 participants who were randomly assigned to the nonrandomized treatment condition were asked to privately select which training session they would prefer to attend and list the reason for their selections. Of these, 131 (62.4%) chose vocabulary training and 79 (37.6%) chose mathematics training. These participants received packets marked "Q" (for quasi-experiment) containing the same posttest materials given to the participants in the randomized experiment, and they were sent to the same training sessions as those who had been randomly assigned to vocabulary or mathematics training. Each training session lasted about 15 minutes. Afterward all participants completed both the mathematics and vocabulary posttests, submitted them to the trainer, and received debriefing. The trainer marked each posttest as to whether the participant had received mathematics or vocabulary training.

3. RESULTS

3.1 Initial Results

Results from the randomized experiment are the presumed best estimate against which all adjusted and unadjusted nonrandomized results are compared. But randomized experiments still encounter group differences in covariates due to sampling error, so we adjusted the randomized results using all of the available covariates in backward stepwise regression. Eventual bias reductions were similar whether we used the adjusted or unadjusted randomized results as a benchmark, however.

3.1.1 The Effects of Mathematics Training on Mathematics Outcome. In the covariance-adjusted randomized experiment, participants who received mathematics training performed 4.01 points (out of 18) better on the mathematics outcome than participants who received vocabulary training (Table 1). In the

Table 1. Percent bias reduction in quasi-experimental results by various adjustments

	Mean difference (standard error)	Absolute bias (Δ)	Percent bias reduction	R^2
Mathematics outcome				
Covariate-adjusted randomized experiment	$4.01_{(.35)}$	.00		.58
Unadjusted quasi-experiment	$5.01_{(.55)}$	1.00		.28
PS stratification	$3.72_{(.57)}$	.29	71%	.37
Plus covariates with strata	$4.05_{(.59)}$	.04	96%	
PS linear ANCOVA	$3.64_{(.46)}$	.37	63%	.34
Plus covariates	$3.65_{(.42)}$	.36	64%	.64
PS nonlinear ANCOVA	$3.60_{(.44)}$	.41	59%	.34
Plus covariates	$3.67_{(.42)}$	.34	66%	.63
PS weighting	$3.67_{(.71)}$	.34	66%	.16
Plus covariates	$3.71_{(.40)}$	.30	70%	.66
PS stratification with predictors of convenience	$4.84_{(.51)}$	.83	17%	.28
Plus covariates	$5.06_{(.51)}$	1.05	−5%*	.35
ANCOVA using observed covariates	$3.85_{(.44)}$	.16	84%	.63
Vocabulary Outcome				
Covariate-adjusted randomized experiment	$8.25_{(.37)}$			.71
Unadjusted quasi-experiment	$9.00_{(.51)}$	.75		.60
PS stratification	$8.15_{(.60)}$	.11	86%	.64
Plus covariates with strata	$8.32_{(.49)}$	.07	91%	
PS linear ANCOVA	$8.07_{(.49)}$	.18	76%	.62
Plus covariates	$8.07_{(.47)}$	.18	76%	.76
PS nonlinear ANCOVA	$8.03_{(.50)}$	.21	72%	.63
Plus covariates	$8.03_{(.48)}$	.22	70%	.77
PS weighting	$8.22_{(.66)}$	.03	96%	.54
Plus covariates	$8.19_{(.51)}$	.07	91%	.76
PS stratification with predictors of convenience	$8.77_{(.48)}$	.52	30%	.62
Plus covariates	$8.68_{(.47)}$	.43	43%	.65
ANCOVA using observed covariates	$8.21_{(.43)}$	.05	94%	.76

NOTE: All estimates are based on regression analyses. Estimated standard errors for propensity score methods are based on 1,000 bootstrap samples (separate samples for each group), with refitted propensity scores and quintiles for each sample (predictors remained unchanged). Each model is presented with only the propensity scores used in the adjustment, and then with the same propensity score adjustment plus the addition of covariates based on backward stepwise inclusion (with main effects only). Standard errors for stratification plus covariates within strata are regression based; bootstrapped standard errors for this adjustment were 68.1 and 28.8 for mathematics and vocabulary outcomes, respectively. The same covariates were directly entered for each stratum adjustment. An overall R^2 cannot be computed for this adjustment.

*This adjustment increased bias by 5%.

unadjusted nonrandomized experiment, the same effect was 5.01 points, or 25% larger than in the randomized experiment. The absolute value of the difference between these results ($\Delta = |4.01 - 5.01| = 1.00$) is a measure of the bias in the unadjusted nonrandomized results, where $\Delta = 0$ indicates no bias.

3.1.2 The Effects of Vocabulary Training on Vocabulary Outcome. In the covariance-adjusted randomized experiment, the participants who received vocabulary training performed 8.25 points (out of 30) better on the vocabulary outcome than the participants who received mathematics training (see Table 1). In the nonrandomized experiment, the same effect was 9.00 points, or 9% larger than in the randomized experiment. The absolute value of the difference between these results is $\Delta = |8.25 - 9.00| = .75$.

3.2 Adjusted Results

There is only borderline evidence indicating that the results from the nonrandomized experiment differ significantly different from those of the randomized experiment. Still, of particular interest in this study is whether the results from the nonrandomized experiment can be made to more closely approximate results from the randomized experiment. We now explore several alternative adjustments to assess the extent to which they offer reductions in the estimated bias.

3.2.1 Using Ordinary Linear Regression. Many researchers would adjust the nonrandomized results using ordinary linear regression, predicting outcome from treatment condition and the observed covariates. This method, with backward selection of main effects only, reduced the estimated bias by 94% for vocabulary outcome and 84% for mathematics outcome. In Table 1, this is the best adjustment for mathematics outcome and the second-best adjustment for vocabulary outcome.

3.2.2 Using Propensity Scores. Although several other kinds of adjustments are possible, such as econometric selection bias modeling (e.g., Heckman et al. 1997), we focus on propensity score analysis because of the transparency of its methods and assumptions, its current popularity, and the ease with which it can be done. For person i ($i = 1, \ldots, N$), let Z_i denote the treatment assignment ($Z_i = 1$ if the person receives treatment, in our study vocabulary training, and $Z_i = 0$ if the person receives no or another treatment, here mathematics training) and let $\mathbf{x}_i$ denote the vector of observed covariates. The propensity score for person i is the conditional probability of receiving the treatment given the vector of observed covariates, $e(\mathbf{x}_i) = \Pr(Z_i = 1|\mathbf{X}_i = \mathbf{x}_i)$, where it is assumed that, given the $\mathbf{X}$'s, the Z_i's are independent. Various authors (e.g., Rosenbaum and Rubin 1983) have shown that methods that equate groups on $e(\mathbf{X})$, like subclassification, weighting, or regression adjustment, tend to produce unbiased estimates of the treatment effects if the assumption of strongly ignorable treatment assignment holds. This is the case if treatment assignment (Z) and the potential outcomes [$Y = (Y_0, Y_1)$, under the control and treatment condition] are conditionally independent given the observed covariates $\mathbf{X}$, that is, $\Pr(Z|\mathbf{X}, Y) = \Pr(Z|\mathbf{X})$, and if $0 < \Pr(e(\mathbf{x}_i)) < 1$, for all $\mathbf{x}_i$. The assumption is met if all variables related to both those outcomes and treatment assignment are included among the covariates (i.e., there is no hidden bias) and if there is a nonzero probability of being assigned to the treatment or comparison group for all persons (Rosenbaum and Rubin 1983).

Using these data, we created propensity scores using logistic regression. All subsequent analyses used logit-transformed propensity scores (Rubin 2001). Correlations between predictors and both choice of condition and outcome variables are given in Table 2. Without looking at the outcome variables, we tried many models for creating propensity scores, selecting the one that maximized balance on Rubin's (2001) criteria: (a) The standardized difference in the mean propensity score in the two groups (B) should be near 0, (b) the ratio of the variance of the propensity score in the two groups (R) should be near 1, and (c) ratio of the variances of the covariates after adjusting for the propensity score must be close to 1, where ratios between .80 and 1.25 are desirable and those $<.50$ or <2.0 are far too extreme. The propensity scores that we used were well balanced using these criteria (Table 3), except that three covariates had variance ratios slightly outside the desirable range (extraversion, 1.357; openness to experience, .799; number of prior math courses, 1.324). They also were well balanced using the criteria proposed by Rosenbaum and Rubin (1984); a 2×5 analysis of variance (treatment conditions by propensity score quintiles) yielded no significant main effect for treatment and no interaction for any of the covariates in this study. Figure 2 presents a kernel density graph of the propensity score logits both for the total sample (with vertical quintile borders) and by condition. Overlap was reasonable except at the extremes, and quintiles all had at least five units in each cell.

Table 1 reports four propensity score adjustments for the nonrandomized experiment: (a) stratification on propensity score quintiles (Rosenbaum and Rubin 1984), (b) use of the propensity score as a covariate in an analysis of covariance (ANCOVA), (c) propensity score ANCOVA including nonlinear (quadratic and cubic) terms, and (d) propensity score weighting (Rubin 2001). Table 1 reports all four adjustments by themselves, and then all four in a model that also includes some of the original covariates entered in a backward-stepwise manner (the rows labeled "Plus covariates"). The table also reports the usual regression-based standard errors, except that most estimated standard errors for methods involving propensity scores were bootstrapped. For each bootstrap sample, the propensity scores were refit; the predictors included remained unchanged.

Overall, the eight propensity score adjustments without covariates reduced bias by an average of 74% (range, 59–96%), depending on the model. Bias reduction was higher for vocabulary outcome ($M = 81\%$; range, 70–96%) than for mathematics outcome ($M = 66\%$; range, 59–73%). Differences in the specific adjustment used were minor and probably should be treated as nonsignificant given the standard errors, although stratification and weighting tended to perform better than ANCOVA. The addition of covariates to any of the propensity score adjustments significantly increased the variance accounted for,

Table 2. Correlations between predictors and outcome in nonrandomized experiments

Predictor	Vocabulary posttest	Mathematics posttest	Chose vocabulary training
Vocabulary pretest	.468**	.109	.169*
Mathematics pretest	.147*	.446**	−.090
Number of prior mathematics courses†	−.018	.299**	−.131
Like mathematics	−.288**	.471**	−.356**
Like literature	.233**	−.226**	.164*
Preferring literature over mathematics	.419**	−.426**	.385**
Extraversion	.005	−.158*	.092
Agreeableness	.120	−.078	.098
Conscientiousness	−.189**	−.041	−.126
Emotionality	−.099	−.115	−.015
Openness to experience	.201**	.050	.053
Mathematics anxiety	−.051	−.140*	.003
Depression†	.087	.149*	−.014
Caucasian	.322**	−.074	.178*
African-American	−.296**	−.015	−.144*
Age†	.077	−.217**	.022
Male	.064	.141*	−.065
Married	−.073	−.162*	.001
Mother education	.094	−.022	.010
Father education	.110	.068	.008
College credit hours†	.132	.125	.033
Math-intensive major	−.169*	.298**	−.191**
ACT comprehensive score	.341**	.418**	.028
High school GPA	−.003	.401**	−.041
College GPA	.059	.219**	−.026

* $P < .05$; ** $P < .01$ (two-tailed).

†These four variables were log-transformed in all analyses to reduce positive skew.

made little difference in bias reduction, and usually but not always reduced the bootstrapped standard errors of the estimate. Estimated standard errors for propensity score weighting were larger than for most other methods, likely inflated by the presence of some very low propensity scores. Standard errors also were high for propensity score stratification, reflecting increased uncertainty about the treatment effect given the coarseness of the strata and the small samples in some cells. Otherwise, estimated standard errors for propensity score–adjusted effects were moderately larger than those for the original covariate-adjusted randomized experiments.

Selecting covariates to use in creating propensity scores is a crucial aspect of good propensity score analysis (Brookhart et al. 2006). The present study was designed to have a rich set of covariates potentially related to treatment choice and outcome. Yet in practice, many researchers create propensity scores from whatever variables are conveniently available. To explore the potential consequences of using only conveniently available covariates, we created a new set of propensity scores using only

Table 3. Rubin's (2001) balance criteria before and after propensity score stratification

	Propensity score		Number of covariates with variance ratio				
Analysis	B	R	≤1/2	>1/2 and ≤4/5	>4/5 and ≤5/4	>5/4 and ≤2	>2
Before any adjustment							
	−1.13	1.51	0	2	17	6	0
After stratification on propensity scores constructed from all covariates							
	−.03	.93	0	1	22	2	0
After stratification on propensity scores constructed from predictors of convenience, balance tested only on the five predictors of convenience							
	−.01	1.10	0	0	5	0	0
After stratification on propensity scores constructed from predictors of convenience, balance tested on all 25 covariates							
	−.01	1.10	0	2	16	7	0

NOTE: Standardized mean difference in propensity scores is given by $B = (\bar{x}_t - \bar{x}_c)/\sqrt{(s_t^2 + s_c^2)/2}$ where $\bar{x}_t$ and $\bar{x}_c$ are the sample means of the propensity scores in the treatment and comparison group, and s_t^2 and s_c^2 the corresponding sample variances. The variance ratio, R, is s_t^2/s_c^2 (also for covariates). Balancing criteria after propensity score stratification are obtained by attaching stratum weights to individual observations (Rubin 2001).

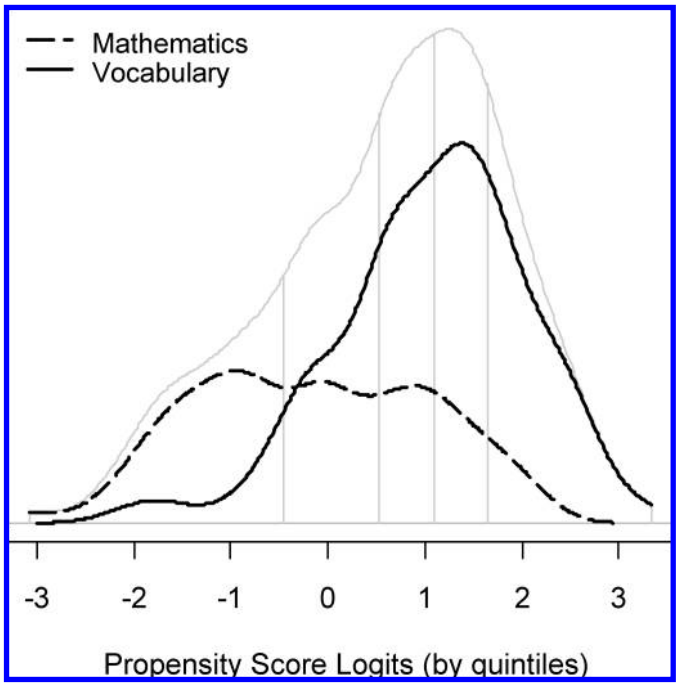

Figure 2. Distribution of propensity score logits smoothed using a kernel density function. The light-gray line represents the total sample, with vertical quintile borders. The dashed line represents those who chose mathematics training, and the solid black line represents those who chose vocabulary training. Negative scores indicate propensity to choose mathematics training.

sex, age, marital status, and race (dummy coded for two predictors, Caucasian and African-American) as predictors. Those variables often are gathered in research and are the kinds of predictors of convenience likely to be available when careful thought has not gone into the inclusion of potential selection variables. Adjusting the results of the nonrandomized experiment by stratifying according to the quintiles of such propensity scores yielded inconsistent, and usually poor, results (Table 1). For the mathematics outcome, this adjustment reduced bias by 17% (and increased bias by 5% when covariates were added); for the vocabulary outcome, this adjustment reduced bias by 30% (43% when covariates were added). Some bias reduction occurred because these four predictors are related to selection (Table 2), but those four predictors clearly are not the only relevant ones.

If a researcher had tested the propensity scores resulting from the five predictors of convenience using Rubin's (2001) balance criteria, they would have performed quite well (Table 3, third row). But this would have hidden a failure to balance well on many of the remaining covariates that presumably would not have been observed by such a researcher (Table 3, fourth row). This is a good illustration of hidden bias and how it might lead to poor estimates of a treatment effect.

4. DISCUSSION

4.1 Adjustments to Nonrandomized Experiments

This study suggests that adjusted results from nonrandomized experiments can approximate results from randomized experiments. This was true for propensity score adjustments, as well as for ordinary linear regression without the use of propensity scores, some implications of which we discuss shortly. All of the recommended adjustments always reduced bias (and never increased it), and did so substantially. Moreover, they did so despite the fact that the nonrandomized study had a small sample size and was not designed to have a well-matched control group before data collection began. These adjustments might have been even better had the study been designed to be larger with a well-matched control group.

The adjustments may have done well in the present case in part because this study is characterized by a very rich set of covariates that are well measured and plausibly related to both the selection process and the outcome measures. Such richness is not always present in data sets for nonrandomized experiments, especially not in those conducted retrospectively. As demonstrated by our analysis of propensity scores based on predictors of convenience, a lack of covariate richness may greatly reduce the accuracy of adjustments. Implicit is a lesson for the prospective design of nonrandomized experiments, that attention to careful measurement of the selection process can be crucial to the success of subsequent analyses.

Furthermore, our experience analyzing this data set suggests that propensity score adjustments may be sensitive to variations in how those scores are constructed. One example is the sensitivity to which covariate balance criteria are used. We found that some propensity scores constructed under Rosenbaum and Rubin's (1984) balance criteria did not meet Rubin's (2001) balance criteria, but those meeting the latter criteria always met the former. The reliance of the criteria of Rosenbaum and Rubin (1984) on significance testing makes it vulnerable to confusing successful balance with low power. The emphasis of Rubin (2001) on the size of the imbalance may be more desirable. Both sets of criteria probably should be reported. We would benefit from further development of ways to create and assess balance (e.g., Imai, King, and Stuart 2008; Sekhon 2007), as well as from better-justified standards for how much balance should be achieved.

The results also were sensitive to how missing data in the predictors were managed. At first, we followed a recommendation of Rosenbaum and Rubin (1984) to create propensity scores separately for groups with different missing-data patterns. But we found that bias reduction was highly sensitive to seemingly minor changes in how those patterns were identified, in one case even increasing bias. Consequently, we moved to more current missing-data methods, but those results also may prove sensitive to which current method is used (D'Agostino and Rubin 2000). In particular, our results might have changed had we used multiple imputation rather than EM-based imputation.

We used logistic regression to construct propensity scores in the present study. Other methods for creating propensity scores exist, including classification trees, boosted regression, random forests, and boosted regression (e.g., Stone et al. 1995; McCaffrey, Ridgeway, and Morral 2004). A simulation conducted by one of our colleagues suggests that propensity score adjustments also may be sensitive to the methods used, and also quite sensitive to sample size (Luellen 2007).

We are currently exploring the sensitivity of the present data set to many of the variations described in the previous paragraphs. Taking them together, however, it may be that the practice of propensity score analysis in applied research may yield adjustments of unknown or highly variable accuracy. This is not surprising for a method as new as propensity score analysis, and points to the need for more clarity about best propensity score practice.

In view of these matters, a pertinent question is why researchers should consider using propensity scores when ordinary linear regression with covariates does as well or better. One situation in which propensity scores could be used is when the design calls for matching treatment and comparison units on a large number of covariates, for example, when constructing a control group matched to an existing treatment group from a large number of potential controls (e.g., Rubin 2001). Without reducing those covariates to a propensity score, the matching process would not be feasible. Another circumstance is when there is uncertainty about assumptions of linearity in ordinary linear regression that stratification on propensity scores might ameliorate. Such exceptions aside, however, in general our results do not support the preferential use of propensity scores over ordinary linear regression for analytic purposes. On the other hand, Rubin (2001) correctly notes that various methods based in propensity scores can be used to better the design of an observational study before the analysis of outcomes is attempted. For example, they can help create comparisons that meet balance criteria, or conversely, identify datasets where adequate covariate balance cannot be achieved. Ordinary linear regression does not lend itself to making such design improvements.

4.2 Comments on the Laboratory Analog Design Used in This Study

Questions may arise about the replicability and generalizability of these results given the design that we used. The design probably is no more labor-intensive than other methods, at least for researchers with access to large research participant pools like those available in university-based introductory psychology classes. Thus testing replication has few obstacles. Minor changes in the method might improve its feasibility and yield. The second author, for example, added a no-treatment control group to this design in a study in progress and also added achievement motivation as an additional predictor of selection. The first author is working to computerize administration of this method, which might allow rapid implementation of more complex assignment mechanisms or allow Web-based implementation to obtain larger sample sizes. We are also creating a version of the study that can be administered over the Internet, allowing us to improve certain features of this study; for example, we can use computer-generated random numbers rather than birth month to do random assignment.

The question of generalization is more serious and has two aspects. The first aspect concerns how the results reported in this study would change over variations of the method that stay within this general laboratory analog paradigm. One could vary the kind of treatment from the current educational one to mimic other substantive areas, such as job training, health, and different aspects of education. Similarly, one could create more time-consuming treatments, although it would be desirable to avoid attrition from both treatment and measurement, because these are separate problems from adjusting for selection into conditions.

A second variation within the laboratory analog method is to study different selection mechanisms, such as cutoff-based assignment mechanisms used in the regression discontinuity design (Shadish et al. 2002), analogs to parental selection of children into interventions, and analogs to the kind of selection that occurs in mental health, where participants choose treatment due to extremely high scores on a latent variable such as distress. Such work could advance an empirical theory of selection for different kinds of treatments, improving the efficacy of adjustments that rely on good prediction of selection.

A third variation of the present method is to explore different design elements or analyses. For example, propensity score matching may benefit when the researcher has a much larger pool of potential control group participants from which to select propensity score matches to a smaller group of treatment group participant scores (Rosenbaum and Rubin 1985; Rubin 2001). This should be easy to test with a variation of the present method. Given that propensity score adjustments also are said to work best in large samples, one could also vary sample size to shed light on sample size requirements and randomly assign proportionately more participants to the nonrandomized experiment. The latter also would decrease the estimated standard errors of adjusted estimates. Similarly, one might examine the effectiveness of additional statistical adjustment procedures, such as econometric selection bias models (e.g., Heckman et al. 1997; Greene 1999).

A fourth variation of our method is to study populations other than introductory psychology students. We used psychology students because we could obtain large numbers of them and could exercise a high degree of experimental control. Other populations can approximate those characteristics, especially if the treatment is short or participation is required. For example, Akins, Hollandsworth, and O'Connell (1982) treated introductory psychology and sociology students solicited for dental fear with a 1-hour, researcher-administered intervention given by audio and videotape in a college laboratory. This could be offered to university or community participants more generally. Aiken, West, Schwalm, Carroll, and Hsiung (1998) used students who were required to take a university remedial writing program to create a study similar to the present one, but without the initial random assignment to assignment method. Such cases may be adapted to remedy the latter lacuna. So might the provision of desirable brief services to community participants, such as stress reduction training, especially if accompanied by payment for participation. One could argue that such examples are not really laboratory analogs anymore—especially if they were also conducted in the community rather than in the laboratory—but if so, so much the better.

The latter observation leads into the second part of the generalization question—whether highly controlled laboratory experiments like the present study yield results that would repli-

cate in research about the effects of longer treatments in settings like the classroom, job training center, or physician office where field experimentation takes place. Some variations on our basic laboratory analog could shed light on this concern, such as the hypothetical medical experiment described in Section 1 at the end of the discussion of doubly randomized preference trials. But attrition from measurement and treatment are prevalent in such applied settings and add additional layers of selection bias that propensity scores were not necessarily designed to adjust, as noted for the study of Janevic et al. (2003) (see also Long et al. 2008). Ultimately, the only way to answer this generalization question is to apply the paradigm in the present study to actual field experiments. Such a study might be hard to sell to funding agencies, especially to problem-focused agencies that might be reluctant to spend extra money to fund the nonrandomized experiment if they are already funding the randomized one. Nonetheless, we suspect that chances to do such studies will present themselves in due course to researchers who are sensitive to the opportunity.

[Received July 2007. Revised December 2007.]

REFERENCES

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., and Hsuing, S. (1998), "Comparison of a Randomized and Two Quasi-Experimental Designs in a Single Outcome Evaluation: Efficacy of a University-Level Remedial Writing Program," *Evaluation Review*, 22, 207–244.

Akins, T., Hollandsworth, J. G., and O'Connell, S. J. (1982), "Visual and Verbal Modes of Information Processing and Their Relation to the Effectiveness of Cognitively-Based Anxiety-Reduction Techniques," *Behaviour Research and Therapy*, 20, 261–268.

Beck, A. T., and Beck, R. W. (1972), "Screening Depressed Patients in Family Practice: A Rapid Technique," *Postgraduate Medicine*, 51, 81–85.

Bloom, H. S., Michalopoulos, C., Hill, C. J., and Lei, Y. (2002), *Can Nonexperimental Comparison Group Methods Match the Findings From a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?*, New York: Manpower Development Research Corp.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006), "Variable Selection for Propensity Score Models," *American Journal of Epidemiology*, 163, 1149–1156.

D'Agostino, R. B., and Rubin, D. B. (2000), "Estimating and Using Propensity Scores With Partially Missing Data," *Journal of the American Statistical Association*, 95, 749–759.

Dehejia, R., and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062.

Dorans, N. J., Lyu, C. F., Pommerich, M., and Houston, W. M. (1997), "Concordance Between ACT Assessment and Recentered SAT I Sum Scores," *College and University*, 73, 24–33.

Drake, C. (1993), "Efects of Misspecification of the Propensity Score on Estmators of Treatment Effect," *Biometrics*, 49, 1231–1236.

Educational Testing Service (1962), "Vocabulary Test II (V-2)," in *Kit of Factor Referenced Cognitive Tests*, Princeton, NJ: Author.

——— (1993), "Arithmetic Aptitude Test (RG-1)," in *Kit of Factor Referenced Cognitive Tests*, Princeton, NJ: Author.

Faust, M. W., Ashcraft, M. H., and Fleck, D. E. (1996), "Mathematics Anxiety Effects in Simple and Complex Addition," *Mathematical Cognition*, 2, 25–62.

Glazerman, S., Levy, D. M., and Myers, D. (2003), "Nonexperimental versus Experimental Estimates of Earnings Impacts," *The Annals of the American Academy of Political and Social Science*, 589, 63–93.

Goldberg, L. R. (1997), "Big-Five Factor Markers Derived From the IPIP Item Pool (Short Scales)," *International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality and Other Individual Differences*, available at *http://ipip.ori.org/ipip/appendixa.htm#AppendixA*.

Greene, W. H. (1999), *Econometric Analysis*, Upper Saddle River, NJ: Prentice-Hall.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997), "Matching as an Econometric Evaluation Estimator: Evidence From Evaluating a Job Training Programme," *Review of Economic Studies*, 64, 605–654.

Heinsman, D. T., and Shadish, W. R. (1996), "Assignment Methods in Experimentation: When Do Nonrandomized Experiments Approximate the Answers From Randomized Experiments?" *Psychological Methods*, 1, 154–169.

Hill, J. L., Reiter, J. P., and Zanutto, E. L. (2004), "A Comparison of Experimental and Observational Data Analyses," in *Applied Bayesian Modeling and Causal Inference From Incomplete Data Perspectives*, eds. A. Gelman and X.-L. Meng, New York: Wiley, pp. 51–60.

Hill, J. L., Rubin, D. B., and Thomas, N. (2000), "The Design of the New York School Choice Scholarship Program Evaluation," in *Validity and Social Experimentation: Donald Campbell's Legacy*, Vol. 1, ed. L. Bickman, Thousand Oaks, CA: Sage, pp. 155–180.

Hirano, K., and Imbens, G. W. (2001), "Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization," *Health Services & Outcomes Research Methodology*, 2, 259–278.

Holland, P. W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–970.

Imai, K., King, G., and Stuart, E. A. (2008), "Misunderstandings Between Experimentalists and Observationalists About Causal Inference," *Journal of the Royal Statistical Society*, Ser. A, 171, 481–502.

Janevic, M. R., Janz, N. K., Lin, X., Pan, W., Sinco, B. R., and Clark, N. M. (2003), "The Role of Choice in Health Education Intervention Trials: A Review and Case Study," *Social Science and Medicine*, 56, 1581–1594.

Kleinmuntz, B. (1960), "Identification of Maladjusted College Students," *Journal of Counseling Psychology*, 7, 209–211.

Lipsey, M. W., and Wilson, D. B. (1993), "The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation From Meta-Analysis," *American Psychologist*, 48, 1181–1209.

Long, Q., Little, R. J., and Lin, X. (2008), "Causal Inference in Hybrid Intervention Trials Involving Treatment Choice," *Journal of the American Statistical Association*, 103, 474–484.

Luellen, J. K. (2007), "A Comparison of Propensity Score Estimation and Adjustment Methods on Simulated Data," unpublished doctoral dissertation, University of Memphis, Dept. of Psychology.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004), "Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies," *Psychological Methods*, 9, 403–425.

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

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

——— (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516–524.

——— (1985), "The Bias Due to Incomplete Matching," *Biometrics*, 41, 103–116.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (2001), "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation," *Health Services and Outcomes Research Methodology*, 2, 169–188.

Rücker, G. (1989), "A Two-Stage Trial Design for Testing Treatment, Self-Selection and Treatment Preference Effects," *Statistics in Medicine*, 8, 477–485.

Sekhon, J. S. (2007), "Alternative Balance Metrics for Bias Reduction in Matching Methods for Causal Inference," Travers Dept. of Political Sciences, Survey Research Center, University of California, Berkeley, available at *http://sekhon.berkeley.edu/papers/SekhonBalanceMetrics.pdf*.

Shadish, W. R. (2000), "The Empirical Program of Quasi-Experimentation," in *Validity and Social Experimentation: Donald Campbell's Legacy*, ed. L. Bickman, Thousand Oaks, CA: Sage, pp. 13–35.

Shadish, W. R., and Ragsdale, K. (1996), "Random versus Nonrandom Assignment in Controlled Experiments: Do You Get the Same Answer?" *Journal of Consulting and Clinical Psychology*, 64, 1290–1305.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston: Houghton-Mifflin.

Stone, R. A., Obrosky, D. S., Singer, D. E., Kapoor, W. N., Fine, M. J., and the Pneumonia Patient Outcomes Research Team (PORT) Investigators (1995), "Propensity Score Adjustment for Pretreatment Differences Between Hospitalized and Ambulatory Patients With Community-Acquired Pneumonia," *Medical Care*, 33, AS56–AS66.

Wennberg, J. E., Barry, M. J., Fowler, F. J., and Mulley, A. (1993), "Outcomes Research, PORTs, and Health Care Reform," *Annals of the New York Academy of Sciences*, 703, 52–62.

Comment

Roderick J. LITTLE, Qi LONG, and Xihong LIN

Randomized designs are the gold standard for experiments that compare treatments, but they exclude individuals not willing to be randomized, as when they have a strong preference for a particular treatment. In behavioral trials that cannot be blinded, it is likely that treatments are most successful for individuals who would choose that treatment if given the choice, because these individuals may be more motivated to comply with treatments they prefer. Thus a design that allows individuals to choose their treatment has the attraction of including individuals who otherwise might not participate in a randomized trial, although most statisticians would say that this advantage is trumped by the virtues of randomization in avoiding selection bias. Hybrid designs that involve randomization and choice arms, like the doubly randomized preference trial (DRPT) applied here, have the potential to combine the advantages of both designs.

The real scientific value of hybrid designs is in assessing not the effects of selection bias, but rather their potential to combine the advantages of randomization and choice to yield insights about the role of treatment preference in a naturalistic setting. Studies with hybrid designs can answer scientific questions that otherwise may not be answered from completely randomized studies.

We appreciate the opportunity to comment on this interesting application of the DRPT design by Shadish, Clark, and Steiner (henceforth SCS) to a psychology class experiment comparing the effects of verbal and math pretest training on posttest scores. The study was carefully designed and provides an opportunity to compare the estimated treatment effects in the randomized and choice groups with and without adjustments for covariates that predict the treatment choice. An attractive feature of the study is the availability of verbal and mathematics outcomes and corresponding pretest variables that measure ability in these domains. Results provide empirical confirmation that adjustment for covariates related to selection and to the outcome can remove, or at least reduce, selection bias.

SCS also compare propensity and multiple regression adjustments, but we feel that their evidence is too limited to allow us to draw any conclusions about this aspect; differences generally are very minor compared with sampling error. Although SCS shed some insights into the comparison of various methods, we think that there are dangers in overgeneralizing from results obtained from a single data set; more comprehensive comparisons based on simulations are needed.

The main goal of the SCS's experiment appears to be to assess the effects of selection bias, rather than to actually compare the effectiveness of the treatments in subgroups with different treatment preferences. Our own work in this area (Long, Little, and Lin 2008) was motivated by the Women Take Pride Study (Janevic et al. 2003), a National Institutes of Health–funded DRPT comparing two behavioral treatments for the management of heart disease in older women, with treatments differing in the mode of administration—a group versus a self-administered format. A standard care control group was also included in the randomization arm. As noted by the authors, this study had issues not present in the study of SCS, in that not all individuals agreed to participate, and there were missing data because of dropouts. On the other hand, this was a real study to assess an actual intervention in the target population of interest, whereas the study of SCS appears to be more of a classroom exercise. The additional problems present in the study of Janevic et al. are "real life" issues encountered by behavioral researchers, and the study of Janevic et al. illustrates that the DRPT design can be successfully implemented and yield meaningful results regarding treatment preference in a real setting.

SCS compare estimates of a treatment (say T) on an outcome (say Y) in the choice and randomization arms of the DRPT study, before and after adjustment for baseline covariates Z. This approach addresses the question of selection bias in the choice arm but does not provide estimates of the effects of treatment preference, which can be estimated from such DRPT designs. Specifically, Long et al. (2008) presented an analysis approach that integrates the information in both the randomization and choice arms to provide estimates of the preference effects on the outcome, using a model founded on the causal inference framework of Rubin (1974, 1978). The key step is to define a preference variable P indicating which treatment an individual would choose (here mathematics or vocabulary training) if given the choice. Because P is a pretreatment variable, it makes sense to condition on P when assessing the effect of T on Y. The value of P equals the treatment received for individuals in the preference arm of the DRPT, and thus is observed, but it is missing for individuals in the randomization arm. Thus the analysis of the effects of P and T on Y (with or without adjustment for Z) can be considered a problem of regression with missing values of the covariate P in the randomization arm. Method-of-moments estimation without covariates was originally described by Rücker (1989). Long et al. (2008) placed Rücker's analysis in a modeling framework and discussed maximum likelihood (ML) estimation methods that also allowed for conditioning on covariates Z.

Two kinds of preference effects can be distinguished: (a) the direct effect of P on Y, which if present leads to selection bias in the estimated effect of T on Y in the choice arm, and (b) the interaction of P and T on Y, which addresses whether treatment preference modifies the effect of treatment on outcome. In

Roderick J. Little is Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: *rlittle@umich.edu*). Qi Long is Assistant Professor, Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322. Xihong Lin is Professor, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115.

Journal of the American Statistical Association
December 2008, Vol. 103, No. 484, Applications and Case Studies
DOI 10.1198/016214508000000995

Table 1. Estimated subpopulation means using an unadjusted ML analysis

Outcome	Preference population	Treatment assignment: Vocabulary training	Treatment assignment: Mathematics training	Treatment effect (SE*; p value)	Preference effect† (SE; p value)
Vocabulary outcome	Vocabulary	16.92	8.18	8.74 (.51;<.001)	2.21 (1.98; .31)
	Mathematics	14.33	7.80	6.52 (1.72; .001)	
Mathematics outcome	Vocabulary	7.42	11.11	−3.69 (.56;<.001)	1.89 (2.24; .42)
	Mathematics	6.58	12.16	−5.58 (1.87; .035)	

*SE and p values computed using a bootstrap method.

†Preference effect is defined as the difference in treatment effect between two preference populations.

Table 2. Estimated subpopulation means using an adjusted ML analysis

Outcome	Preference population	Treatment assignment: Vocabulary training	Treatment assignment: Mathematics training	Treatment effect (SE; p value)	Preference effect (SE; p value)
Vocabulary outcome	Vocabulary	16.74	8.44	8.29 (.52;<.001)	−.02 (2.34; .98)
	Mathematics	16.06	7.74	8.32 (2.06;<.001)	
Mathematics outcome	Vocabulary	7.36	11.40	−4.05 (.43;<.001)	−.076 (1.29; .93)
	Mathematics	8.26	12.24	−3.97 (1.06; .002)	

Table 3. Estimated subpopulation means from additive model for training and preference on outcome, unadjusted ML analysis

Outcome	Preference population	Treatment assignment: Vocabulary training	Treatment assignment: Mathematics training	Treatment effect (SE; p value)
Vocabulary outcome	Vocabulary	16.62	8.42	8.20 (.41;<.001)
	Mathematics	15.86	7.66	8.20 (.41;<.001)
Mathematics outcome	Vocabulary	7.14	11.28	−4.14 (.40;<.001)
	Mathematics	7.91	12.05	−4.14 (.40;<.001)

Table 4. Estimated subpopulation means from additive model for training and preference on outcome, adjusted ML analysis

Outcome	Preference population	Treatment assignment: Vocabulary training	Treatment assignment: Mathematics training	Treatment effect (SE; p value)
Vocabulary outcome	Vocabulary	16.74	8.44	8.30 (.39;<.001)
	Mathematics	16.04	7.74	8.30 (.39;<.001)
Mathematics outcome	Vocabulary	7.36	11.39	−4.03 (.33;<.001)
	Mathematics	8.22	12.24	−4.03 (.33;<.001)

the setting of SCS, consider, for example, the mathematics outcome. Question (a) concerns whether those who prefer mathematics training have a different (one would surmise lower) mathematics outcome than those who prefer vocabulary training. Question (b) concerns whether the advantage of mathematics training over vocabulary training on mathematics outcome is different (one would surmise greater) for those who would choose mathematics training ($P = M$) and those who would choose vocabulary training ($P = V$). These questions can also be considered after conditioning on covariates Z. Parallel questions can be formulated for the vocabulary outcome.

We applied the methods of Long et al. (2008) to SCS's data, and present key results in Tables 1–4. (More details are available on request.) Tables 1 and 2 show the effects of T and P on the outcome based on models that include the interaction between T and P. Table 1 is unadjusted, and Table 2 is covariate-adjusted using the models of SCS. More specifically, the models for the primary outcomes of interest are based on covariates used in SCS's analysis for their randomized experiment, and the propensity score model is based on all covariates. The same transformations are used for a subset of covariates. The estimates are computed by ML; similar results for the unadjusted analysis are obtained by the method of moments (Long et al. 2008).

Table 1 shows that for the vocabulary outcome, the effect of vocabulary training relative to mathematics training is esti-

mated to be 8.74 in the subpopulation who prefer vocabulary training and 6.52 in the subpopulation who prefer mathematics training. Thus there is a modest preference effect of 2.21 in the expected direction, although this is not statistically significant compared with its associated standard error (SE) of 1.98. For the mathematics outcome, the effect of mathematics training relative to vocabulary training is 5.58 in the subpopulation that prefers mathematics training and 3.69 in the subpopulation that prefers vocabulary training. Thus there is a small preference effect of 1.89, again in the expected direction, but not statistically significant. Table 2 shows that adjustment for covariates has the effect of reducing the preference effects to values close to 0, suggesting that the covariates are accounting for the preferences effects. This parallels the role of the covariates in reducing the effects of selection bias due to treatment choice in the SCS analysis.

Because the preference effects are minimal, we also fitted a model that assumes that the effects of training and preference are additive. Results from these models are given in Table 3 for an unadjusted analysis and in Table 4 for an adjusted analysis. From these results, we see that the main effects of preference are very small compared with the main effects of training, and are not statistically significant. The effects of training from the additive model are close to SCS's estimates from the randomized arm of the experiment, indicating that our model also removes the effects of selection bias in the preference arm. One possible reason for the modest effects of preference here is that this is a "low-stakes" setting, because presumably students are not being rewarded or penalized based on their performance.

There were missing values in SCS's data and our analysis was conducted based on the imputed data set used in their analysis. This has the limitation of ignoring imputation uncertainty. Because our analysis treats values of P as missing, it can be quite easily extended to handle missing values in other variables, using the EM algorithm or Bayesian simulation methods.

Long et al. (2008) noted two key assumptions underlying these analyses of DRPT's. The exclusion restriction (ER) assumption states that the outcome for an individual under a particular treatment is the same whether that individual is randomized to that treatment or chooses that treatment. The no selection bias from randomization (NSBR) assumption states that the randomization and choice arms of the DRPT are samples from the same population. Both the ER and NSBR assumptions seem reasonable in the relatively controlled conditions of SCS's study but may be questionable in other applications. In particular, the NSBR assumption is questionable when there are individuals who would participate in a study if allowed to choose their treatment but would not if they were randomized to treatment. In their section 7, Long et al. (2008) discussed generalizations of DRPT's that allow identification of parameters when these assumptions are relaxed. This material might serve as a useful framework for additional work on hybrid trials, which we believe to be a fruitful area for future study design.

ADDITIONAL REFERENCE

Rubin, D.B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58.

Comment

Jennifer HILL

I would like to thank the authors for their thought-provoking and genuinely useful article on what I feel to be a crucially important topic. I am always pleased to see work that investigates the capacity for observational studies to produce unbiased treatment effect (causal) estimates using a thoughtful approach, and this article certainly is one of the best conceptualized and operationalized efforts in this genre. I have no direct criticisms of the proposed design in itself. I instead discuss this contribution within a broader framework of approaches that can be used to evaluate the efficacy of the types of quasi-experimental and observational studies often performed to explore the trade-offs that exist. To avoid confusion regarding the distinction between typical research studies and the class of research used to evaluate their effectiveness at answering causal questions, for the remainder of this discussion I refer to any study in the latter group as a "confirmatory evaluation study" (CES).

There is a (slowly) growing literature on the ability of observational studies to estimate causal effects. But given the overwhelming number of applied studies in the social sciences, public health, and other fields that use quasi-experimental or observational designs to draw causal conclusions (and the fact that these rely on untestable assumptions), it is somewhat shocking to see how little work has been done to determine how likely it is and under what circumstances such endeavors can be expected to actually yield reliable answers. This has to do in large part with a lack of deep understanding regarding what it takes to satisfy ignorability in practice. In some disciplines, unfortunately, ignorability is assumed without a second thought. In others, almost equally dangerously, ignorability of the treatment assignment is dismissed out of hand as unachievable in the absence of an experiment (or sometimes a "natural experiment," although these studies typically arrive with their own baggage). Neither perspective motivates research to deepen our understanding about the conditions under which we just might be able to use observational data to draw causal conclusions—information vital to applied researchers.

J. Hill is Professor, Steinhardt School, New York University, New York, NY 10003 (E-mail: *jennifer.hill@nyu.edu*).

Journal of the American Statistical Association
December 2008, Vol. 103, No. 484, Applications and Case Studies
DOI 10.1198/016214508000001002

Shadish, Clark, and Steiner (henceforth SCS) present their proposed CES as the ideal form in this genre. Although I understand their point, and appreciate the strengths of their design and its advantages over alternatives, I still see it as one of several different options, each of which addresses different goals with more or less success depending on the difficulty of the research question. My discussion places their CES within a broader context to highlight the situations in which it is superior versus those that are best addressed by other options.

THE SCS PROPOSAL

The design that SCS advocate could be exceptionally useful for evaluating the potential effectiveness of a certain class of research studies. It is most obviously applicable to studies that have a treatment of high intensity (i.e., where we would expect strong treatment effects, as in this study) but short duration, a short time between pretest (where I use pretest to generically refer to the pretreatment measure of the subsequent outcome of interest) and posttest and between treatment and posttest. Also, there should be control over confounders that are measured, the availability of many highly informative potential confounders, a relatively small amount of missing data, and an absence of noncompliance with treatment assignment. Their design also is most directly applicable to studies performed in a laboratory setting with relatively easy-to-recruit participants (given that their time commitment is low and they typically can be rewarded for participating in the study).

SCS are forthright about the fact that their proposal has limitations with regard to generalizability. As a partial remedy, they describe variations of it that would allow the results to inform a broader class of studies. For instance, they propose using it with different kinds of treatments, more time-consuming treatments, and study populations broader than simply psychology students at a university (e.g., students from other majors or even perhaps members of the surrounding community). They also suggest studying different selection mechanisms and varying the sample size. Each of these variations is plausible and could indeed expand the scope in useful directions.

Unfortunately, these variations do not go nearly far enough toward informing the vast majority of intervention studies, let alone purely observational studies, that are typically performed. I first enumerate several reasons for this and then discuss how alternate forms of CESs can address these issues.

Problems of Expense. The vast majority of interventions that I encounter in my research have treatments (programs, policies, medical treatments) that last for weeks, months, or even years. Follow-up can take place over many years as well. Neither the program nor the evaluation takes place in a laboratory setting. Recruiting of participants is not particularly easy or cheap. SCS address this to a certain extent by saying that in theory, their proposed CES design could be applied to field experiments (which presumably could have all of these features). They acknowledge, however, that it might be difficult to convince funders to pay for the added study. Given that it is often difficult to secure funding for sample sizes sufficient for estimating anything but main effects in these studies, I think it could be extraordinarily difficult to secure the necessary additional funding for the SCS design. I do wholeheartedly agree, however, that researchers should be looking for opportunities to make this happen. In the meantime, we still need to get some answers.

Randomized Experiments Are a Must. A necessary feature of this design is that the intervention must be able to be evaluated by a randomized experiment. This precludes study of designs for interventions that cannot be studied in this way due to legal, ethical, or logistical reasons. Ironically, of course, these are the causal questions for which we most need to find reliable nonexperimental solutions!

Going Beyond the Average Treatment Effect. SCS's design (as proposed) is not useful for estimating such effects as the effect of the treatment on the treated, which often can be more useful from a practical or policy perspective.

Mapping to Less Pristine Studies Is Unclear. Many observational studies do not have the luxury of having a control group recruited at the same time as the treatment participants and to whom the same survey instrument was administered. They often have higher rates of missing data. Invention studies often suffer from noncompliance. Moreover, many observational studies do not have interventions that actually have been manipulated (although most would argue that these "causes" must be manipulable in theory). It is easy to dismiss such studies as not worth performing; however, this would be unrealistic, because the vast majority of studies that try to answer causal questions probably suffer from at least one, if not several, of these issues. At the very least, these less pristine studies can be used as (typically substantially less expensive) pilot studies to inform future randomized studies, so the more reliable the information that comes out of them, the better designed the more rigorous follow-up studies can be.

Because of these disconnects, the SCS proposal is not well calibrated to the messier settings and questions that comprise the vast majority of questions posed in social science and public policy (my intellectual home). Therefore, I describe other options, revisiting some of options that SCS discussed briefly and introducing some others.

SIMULATIONS

SCS's discussion of simulations undersells the potential usefulness of this approach because of an overly narrow definition. Simulations need not be so divorced from the real world as CES describe (though they certainly can be, and that is something to watch out for). One useful and increasingly common approach is to construct a simulation that uses data observed in an actual study (see, e.g., Imai and van Dyk 2004; Hill and McCulloch 2008). For instance, the covariates and the treatment assignment indicator from a real observational study or quasi-experiment could be used outright. Then the only "fake" data to be generated would be the (potential) outcomes. These could be simulated under varying assumptions about the parametric form of the response surface (e.g., linear, normal errors, constant or heterogeneous treatment effects) and ignorability. In fact, a sensitivity analysis could be built in directly by generating "unobserved covariates" with varying levels of association with treatment and outcome.

Simulations address all four of my concerns raised earlier. They can be used in any setting, even with retrospective observational data, and for any estimand. They can even directly use

the data from the specific study that one would like to inform. Simulations are a much lower-cost option; they do not require additional sample size study for an intervention study or even the gathering of new observational data. The fact that parametric forms must be specified for certain variables is a drawback in some regards; however, it allows for insight into a wider range of studies. Another drawback is that simulations cannot directly address the ignorability issue, however, they can address it indirectly through sensitivity analyses.

CONSTRUCTED OBSERVATIONAL STUDIES

I use the term "constructed observational studies" to refer to what SCS call the "single-study" approach because it is more evocative of what is actually being created. (As a side note, these studies sometimes combine elements from more than one study so the term "single-study" might be misleading.) Constructed observational studies are created using randomized experiments, and thus also automatically suffer from the second weakness described earlier. SCS rightly point out that these studies also can suffer from the weakness that they "confound assignment method with other study features," which may make it more complicated to figure out what went wrong when the estimates do not line up (though see below for examples of constructed observational studies for which these criticisms are less valid). However I would argue that these weaknesses can often be offset by other strengths.

First, at a practical level, constructed observational studies may be a more reasonable alternative logistically and financially, thus addressing the first concern. But this is not just an issue of convenience. Constructed observational studies have the capacity to create a closer mapping between the CES and the types of studies that are capable of being carried out; this addresses some of the issues raised earlier. I illustrate with a few examples.

The first constructed observational study performed (to my knowledge) used data from a large field experiment called National Supported Work (LaLonde 1986). Here the treatment was a job training program that targeted disadvantaged young men. This program lasted 12 to 18 months, and follow-up took place over several years. Nonexperimental control groups were constructed using the PSID and CPS, an approach that SCS criticize due to the lack of connection with the original study; for instance, the controls were pulled from different locations and had different survey instruments. Clearly, this is not an ideal observational study design. But one of the motivations for this design was that it reflected (and thus was highly relevant for) the types of comparison group observational studies being used to evaluate job training (and similar) programs at the time. This study is becoming a classic, and the data have been used in several studies to explore the effectiveness of a range of different methodological approaches (Dehejia and Wahba 1999; Diamond and Sekhon 2008; Hill and McCulloch 2008).

The Infant Health and Development Program (IHDP) was an intervention that targeted low birth weight and premature children and provided them with such services as home visits and intensive child care in the first few years of life. Follow-up assessments were administered at treatment end (age 3) as well as 2, 4, and even 15 years posttreatment. Because the program started at birth, no pretests (i.e., pretreatment versions of the cognitive outcome measures) were available. This study has been used to form different kinds of constructed observational studies (Hill, Reiter, and Zanutto 2004) and simulations (Hill and McCulloch 2008). But one such CES avoids some of the pitfalls highlighted by SCS because it uses only IHDP participants (no other observational data are used) who thus were subject to the same eligibility criteria, general study conditions and time frame, and even the same survey and testing instruments. The trick used here to construct the observational study was to delete a nonrandom portion of the treatment group (in this case, children with African-American mothers) and delete the ethnicity identifier. This constructed study can be used to try to identify the treatment effect either for the remaining treatment children (using the full control group as a comparison group) or for the full control group (using the remaining treatment children as the comparison group). This strategy also has the advantage of being able to be used to replicate either a situation in which there is known to be full overlap in the covariate distributions, because of the initial randomization (when estimating the treatment effect for the remaining treated children) or the reverse (when estimating the treatment effect for the full control group). It has the disadvantage that the "treatment assignment mechanism" that has been created does not map to the true treatment intervention.

Another strategy that can work entirely within the confines of one randomized experiment was performed using a large multisite, multiperiod randomized evaluation of welfare-to-work programs (Bloom, Michalopoulos, Hill, and Lei 2002). In this case researchers attempted to identify treatment effects for treated cohorts (defined by time and location) using control cohorts from other time periods and locations. This design may be more effective in situations where the program and outcomes of interest (here earnings) are not themselves so location- (i.e., labor market) and time-sensitive.

An example of a constructed observational design variant that directly targets the effect of the treatment on those who chose to participate can be found in a study that makes use of a randomized "get-out-the-vote" field experiment (Arceneaux, Gerber, and Green 2006). The constraints that existed in the initial randomized design (i.e., treatment implementation that eliminated the possibility of access among the controls and made the exclusion restriction more plausible) provided evidence that an instrumental variables analysis could yield a valid estimate of the effect of the treatment on those who chose to participate. This particular study was weakened by an insufficiently strong set of measured confounders to support ignorability, but the design could be useful in future studies. Given that this design also requires identification of an instrumental variables estimate, it should be used only in highly controlled experimental settings. For instance, Piekes, Moreno, and Orzol (2008) used a similar design, but in a far less pristine situation where exclusion was not assured, making it difficult to know how to interpret the results.

VARIATIONS ON THE DRPT THEME

One topic that SCS do not address is the potential for interest in estimands other than the average treatment effect (ATE), $E[Y(1) - Y(0)]$. Consider the situation in which we are most

interested in identifying the effects of an intervention on people who will voluntarily elect to participate in it. Here the estimate of interest is the effect of the treatment on the treated, $E[Y(1) - Y(0)|Z = 1]$, which is equal to the ATE only if treatment effects are additive (or in degenerate cases in which there is perfect bias cancellation).

The CES that SCS propose does not directly inform this question. An alternate but related design could be created that would do this. First, ask the CES participants to choose a treatment (with the knowledge, for ethical purposes, that there are a limited number of slots and so not everyone will be able to get their desired treatment). Then a subset of those who chose the treatment could be used to create a randomized study (within which a random proportion of the participants would not be given access to the treatment) and the rest of those who chose the treatment, along with those who did not choose the treatment, would remain in the quasi-experiment. This might require larger sample sizes than needed for the SCS design.

Critics could argue that participating in an experiment in which one is told up front to make a choice that may not be fulfilled may create a different experience, and thus a different treatment effect, than simply choosing to participate in a program or not. This is possible, although many programs have waiting list, so this is not such an unnatural situation. Moreover, the CES that SCS propose is not immune to such arguments; for instance, there might be a difference in the CES that SCS advocate in the motivation levels between those who were assigned to their treatment and those who were given their choice. Indeed, by definition, it will always be impossible to completely replicate the observational design of interest in the form of a randomized experiment, so we have to pick our battles.

RECONCILING A DISTINCT OBSERVATIONAL STUDY AND RANDOMIZED EXPERIMENT

Another potentially useful exercise is to try to reconcile the differential findings of a completely distinct observational study and randomized experiment. I provide a somewhat simplified description of a current example that has sparked a good deal of controversy. (For a more detailed description, see Hernan et al. 2008.) This research examines the substantial and important differences between the effects of estrogen-progestin hormone replacement therapy (HRT) on coronary heart disease estimated in two different high-profile studies. The Nurses Health Study (NHS), an observational study, found beneficial effects of HRT on coronary heart disease; however, the Women's Health Initiative (WHI), a randomized experiment designed to investigate this problem, found harmful effects. This example has been used for the past few years as a cautionary tale for why we should not trust the results from observational studies. (See, e.g., Taubes 2007, where it was the key motivating anecdote.)

But researchers who looked more closely at these two studies found several important differences between them (Hernan et al. 2008). For instance, the populations in each study were noticeably different in several important respects. As an example, NHS participants at baseline had experienced menopause more recently than the WHI participants at time of hormone initiation. Equally important, there was a crucial difference in what was being estimated. The NHS made comparisons between current HRT users and never users. (This can be considered a form of treatment on treated analysis.) The WHI, on the other hand, randomly assigned postmenopausal women to HRT or placebo and then made comparisons between these groups *as assigned*, regardless of whether those assigned to the HRT arm continued to receive the treatment or whether those assigned to the non-HRT continued to abstain. In other words, in the WHI, an intention-to-treat analysis was performed.

To create comparability, the researchers limited the NHS sample to those who initiated use of HRT at baseline and those with no hormone use at baseline. Other differences between the groups were adjusted for through inverse probability of treatment weighting. Once these differences were accounted for, the results between the two studies were quite similar (Hernan et al. 2008). As in constructed observational studies (of which this might be viewed as a special case), we cannot know whether convergence of estimates in scenarios such as this necessarily implies that the observational study was sufficient to estimate causal effects (or that a similar one would be in the future), but the results are suggestive.

The reverse move toward comparability also could have been performed (and indeed in subsequent work that has yet to be published, this same research team has done just that). That is, one could try to create comparability between the adherence-adjusted effect from the randomized experiment and the effect of the treatment on the treated in the observational study.

META–ANALYSIS

SCS may have oversold meta-analysis. At one logical extreme, surely combining the results of 100 poorly done studies will only lead one to be overconfident in the wrong answers that were obtained. But even if each of these studies were extremely well done and could be counted on to estimate a causal effect, each would likely be estimating causal effects for a different population. Thus, even combining results from 100 well-done studies may not help elucidate much of anything.

MISSING DATA AND NONCOMPLIANCE

As a final note on missing data and noncompliance, SCS promote the full participation rate and near lack of outcome missing data as strengths of this study and advocate for mimicking these features. But this is not a realistic goal for the vast majority of studies that exist. Researchers and program evaluation firms have been struggling for years to decrease missing data rates in studies, but it is practically impossible to get even 90% response rates over time. Regarding noncompliance, from a practical perspective, it is unclear that we want to encourage or force full compliance in these sorts of studies unless the "real-world" analog (program or policy) similarly will force compliance.

CONCLUSION

In conclusion, I want to emphasize my respect for this article's focus on using both a reasonable observational study and an extremely well-calibrated CES to evaluate its usefulness. In contrast to SCS's approach, sometimes it seems that researchers are so concerned with not appearing to overly advantage a particular observational method that they neglect to make reasonable choices regarding how to perform the observational study and thus doom it from the beginning. Although I understand

the need to enter such projects with a good deal of healthy skepticism about the capacity of observational studies, failing to struggle to make the observational study work does not help us figure out under what circumstances we might actually be able to get one to work. Full disclosure always can be achieved by apprising the reader about the paths that did not succeed.

The debate over the article of Dehejia and Wahba (1999, henceforth DW) using the National Supported Work (NSW) constructed observational study is emblematic of this struggle. The initial work was positive toward propensity scores approaches. The article was severely criticized, but most (if not all) of the arguments rested on analysis choices that would not be supported by the propensity score literature (e.g., choosing a propensity score model without checking balance on the sample for which it was used) or that made ignorability implausible. With regard to the latter, a primary criticism of the article was that DW used only a select sample of the NSW data set within which propensity score–based treatment effect estimates are much more stable (Smith and Todd 2005). However, DW chose this sample because of the well-known phenomenon in labor economics (sometimes referred to as the "Ashenfelter dip") that individuals who participate in job training programs tend to experience a "dip" in earnings just before entering, and thus it is not sufficient to control for only one period of pretreatment earnings when evaluating job training programs (Ashenfelter 1978). Because DW are labor economists, they took this advice seriously and restricted their sample to those participants for whom two periods of pretreatment earnings were observed. By following good practice, they ended up with a sample that indeed is fairly insensitive to propensity score specification. As punishment for following good practice, they were accused of trying to unfairly favor propensity score matching. A positive outcome of the debate, however, is that in a reply, Dehejia (2005) provided many new analyses (including on the broader data set) and more fully explored when we might expect the method to break down.

A related point is that it would be helpful when methods appear to "succeed" that the authors try to determine what conditions appear to have been most important for that outcome. SCS's design largely eliminates explanations beyond a failure to satisfy ignorability. But within that realm, they explore choices of covariate sets (though arguably they could have done more on this front). Moreover, when a CES demonstrates that an observational study "fails," it would be helpful if rather than simply being sold as a "cautionary tale," the researchers were to fully explore where the breakdown may have occurred. In this regard, the CES that SCS propose narrows the range of choices, which is a definite strength of the approach

I hope that SCS's article motivates equally compelling additional research in this area.

ADDITIONAL REFERENCES

Arceneaux, K., Gerber, A. S., and Green, D. P. (2006), "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment," *Political Analysis*, 14, 37–62.

Ashenfelter, O. (1978), "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, 6, 47–57.

Dehejia, R. (2005), "Practical Propensity Score Matching: A Reply to Smith and Todd," *Journal of Econometrics*, 125, 355–364.

Diamond, A., and Sekhon, J. (2008), "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies," technical report, U.C. Berkeley, Dept. Political Science.

Hernan, M., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Willett, W. C., Manson, J. E., and Robins, J. M. (2008), "Observational Studies Analyzed Like Randomized Experiments: An Application to Postmenopausal Hormone Therapy and Coronary Heart Disease," *Epidemiology*, 19, 766–779.

Hill, J. L., and McCulloch, R. E. (2008), "Bayesian Nonparametric Modeling for Causal Inference," technical report, New York University, Steinhardt, Dept. Humanities and Social Sciences.

Hill, J., and Reiter, J. (2006), "Interval Estimation for Treatment Effects Using Propensity Score Matching," *Statistics in Medicine*, 25, 2230–2256.

Imai, K., and van Dyk, D. (2004), "Causal Inference With General Treatment Regimes: Generalizing the Propensity Score," *Journal of the American Statistical Association*, 99, 854–866.

LaLonde, R. (1986), "Evalauting the Evalautions of Training Programs," *American Economic Review*, 76, 604–620.

Piekes, D. N., Moreno, L., and Orzol, S. M. (2008), "Propensity Score Matching: A Note of Caution for Evaluators of Social Programs," *The American Statistician*, 62, 222–231.

Smith, J., and Todd, P. (2005), "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, 125, 305–353.

Taubes, G. (2007), "Do We Really Know What Makes Us Healthy?" *New York Times Magazine*, September 16.

Comment: The Design and Analysis of Gold Standard Randomized Experiments

Donald B. RUBIN

To start, I must congratulate the authors of Shadish, Clark, and Steiner (henceforth SCS) for contributing this very stimulating article for discussion. The article is important for the ideas that it presents, but even more important for the implicit questions that it raises but does not actually address. Here I focus on only one such question, one that I think other discussants will not address but has broad implications: How should we design and analyze randomized experiments whose answers will be used as gold standards for assessing competitive methods of

Donald B. Rubin is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: *rubin@stat.harvard.edu*). Thanks are due to Cassandra Wolos for performing the repeated propensity score estimations and diagnostics that are referenced in this discussion. Thanks also are due to Elizabeth Zell for her repeated careful readings of the discussion and very helpful comments. This work was supported in part by National Science Foundation grant SES 0550887 and by National Institutes of Health grant R01 DA023879-01.

Journal of the American Statistical Association
December 2008, Vol. 103, No. 484, Applications and Case Studies
DOI 10.1198/016214508000001011

dealing with data from presumably parallel observational studies, such as the doubly randomized preference trial (DRPT) of SCS? I think that the current common practice is not generally good enough.

1. COVARIATES AND THEIR USE IN GOLD STANDARD EXPERIMENTS

Gold standard randomized experiments must be designed to collect extensive covariate information, X, for two principal reasons. First, the X data should include information that could be used to help make treatment decisions in parallel observational studies, because these X's (if plausibly related to the outcomes) must be available to ensure the approximate ignorability of the assignment mechanisms in the observational studies; more on this point in Section 6. Second, the X data should include covariates thought to be prognostically related to important outcomes, to ensure that random imbalances in their distributions have not distorted the results of the randomized experiments. Of course, we rarely can be entirely confident in practice that the collected covariates fully satisfy these criteria, but when designing gold standard experiments, we should try to be so.

SCS included an extensive collection of covariates, but they did not use them in design as effectively as they might have, in either the initial random assignment of units into the randomized experiment versus the observational study or the subsequent random assignment into the mathematics arm or the vocabulary arm. Evidentially, SCS essentially relied on complete randomization (with some blocking in time) to balance covariates, rather than using randomized blocks or some other design technique for increasing precision, that is, for reducing conditional biases due to random covariate imbalances. For gold standard answers, complete randomization may not be good enough, except for point estimation in very large experiments.

2. A RECREATED DBR, WGC, RAF STORY CONCERNING RERANDOMIZING

This issue stimulates an old memory of a question that I asked my PhD advisor, Bill Cochran (WGC), in the late 1960s: What was his advice if, in a randomized experiment, the chosen randomized allocation exhibited substantial imbalance on a prognostically important baseline covariate? According to my memory, his reply had two layers: (1) WGC: Why didn't you block on that variable?; DBR: Well, there were many baseline covariates, and the correct blocking wasn't obvious; and I was lazy at that time; and (2) WGC: This is a question that I (WGC) once asked Fisher (RAF), and RAF's reply was unequivocal. Recreated (hearsay via WGC) RAF: Of course, if the experiment had not been started, I would rerandomize. My memory is that RAF's reported (via WGC) answer continued with something like the following: The standard analysis of the resulting data would understate the true precision, but would not be misleading, especially if one applied ANCOVA with the offending covariate to analyze the rerandomized data set.

I still believe that this old advice is sound. If a randomized allocation likely will lead to an imprecise result (i.e., the allocation is one with substantial potential for conditional bias given the observed values of covariates), then one should rerandomize. Of course, with many baseline covariates, this implicit advice was difficult to implement decades ago, because it effectively called for analyses of many covariates, their interactions, and other nonlinear terms, but it has become much easier to implement with currently available computational power. Because no outcome data are available at this stage, it is impossible to bias intentionally the subsequent results toward any conclusion by such rerandomizations that achieve better multivariate balance on prognostically plausible covariates.

My view is that if diagnostics, such as those based on propensity score analyses, suggest that important imbalances exist, rerandomize, and continue to do so until satisfied, and record reasons for discarded randomizations. If the covariates are numerous relative to the size of samples, and especially if certain covariate values are rare, then judgment will be needed to assess which trade-offs should be made. For instance, large imbalances in prognostically dubious covariates, such as higher-order interactions that are illogical, should be judged as less important than smaller imbalances in prognostically important covariates.

Such random imbalances do occur with the SCS data. For example, the t statistic for baseline "college grade point average (CGPA)" between the experimental and observational study arms approaches -2.6, those for "like math" and "number of math courses" are both more negative than -2, and for the interaction of "agreeableness" and CGPA, it is nearly -3.3. Moreover, for the rare event of being married, the proportion in the observational study arm is nearly 8%, more than twice that in the experimental arm (only 3%). Of course, these large differences were found after the examination of many covariate differences, but their existence cannot be denied. Would we have been more comfortable when comparing the answers from the experimental and observational study arms if these baseline differences had not been present because they were controlled in the design? I think that the answer is "yes."

For more examples of such random imbalances, in the SCS randomized arm there are more than twice the proportion of males with mathematics-intensive majors in the vocabulary arm (14/116; 12.1%) than in the mathematics arm (7/119; 5.9%). Does this suggest that the vocabulary arm has the potential to show more positive estimated effects on the math tests relative to the mathematics arm? Would we have been more comfortable with a different randomization, especially because (not surprisingly) various covariates and interactions have t-statistics around 2 in absolute value (e.g., the vocabulary arm appears more emotional at baseline than the mathematics arm)? Again, I think that the answers are "yes."

3. DESIGN AFTER ASSIGNMENTS

Now suppose that we have completed random assignments for all units and that we do not have the opportunity for altering these assignments. Should we attempt through design (i.e., without access to any outcome data) to control for observed random imbalances in covariate distributions between treatment groups, such as observed in the SCS data? I believe that the answer to this question is also "yes." My view is that when striving for gold standard answers, either because they must play that role for comparison with other answers (as in SCS) or because they are regarded as providing the definitive answers to important questions (as, e.g., with planned definitive medical trials), such postrandomization design efforts

are nearly always wise. That is, we should be convinced that gold standard answers would not be materially altered had the trial been either a randomized block that ensured balance on prognostically important covariates or a rerandomized design that avoided such chance imbalances, rather than a completely randomized trial. The only current general methodology for achieving such balance with many covariates through design is the class of propensity score techniques—particularly, in this situation, repeated propensity score subclassifications, followed by diagnostic assessments. I strongly advocate their use for the postassignment (but pre-outcome data) design of those randomized trials to be used for creating gold standard answers.

4. ATTENDANT ISSUES

Of course, many issues are associated with the application of such methods. Critical among these is the choice of diagnostic techniques to use to assess overall balance among treatment groups. Previously, Paul Rosenbaum and I offered some diagnostic advice for use with subclassification (Rosenbaum and Rubin 1984), and I offered some other advice (Rubin 2001) for matching and subclassification. SCS use both types of diagnostics. Different, but generally consistent methods, are suggested by Imbens and Rubin (2008) for matching and subclassification. Other diagnostics (Wolos and Rubin 2008) involve "Love" plots (Ahmed et al. 2006) in the context of a large randomized trial in South Africa serving a gold standard for a medical intervention (Zell et al. 2007). As noted earlier, scientific judgment will always be needed to assess trade-offs, and having a variety of outcome-free diagnostics available to help assess balance is obviously beneficial. In contrast to the focus here on diagnostics in the design phase, most of the diagnostic assessments of SCS are based on outcome analyses.

Another attendant issues is whether to restrict inferences in the randomized experiment to subgroups of units that are represented in both randomized arms, if the propensity score analyses and diagnostics reveal that there are units in one arm that are unrepresented in the other. In the SCS database, it seems that only relatively contrived definitions of imbalance lead to the conclusion of lack of adequate overlap in either of the randomized experiments.

5. ANALYSES IN CONDITIONALLY BALANCED RANDOMIZED EXPERIMENTS

Suppose that, by design, in the senses conveyed in Sections 2, 3, and 4, we have achieved balanced treatment arms with respect to multivariate X, either overall (in the sense of Sec. 2) or within subclasses and strata (in the sense of Secs. 3 and 4), better balance than typically could be achieved by complete randomization. For simplicity, I call this a conditionally balanced experiment. Only now we can examine outcome data.

Should model-based adjustments for X (e.g., ANCOVA) be applied? Such adjustments, which should be carried out within the balanced groups separately (e.g., as referenced in SCS's table 1, row 4 for mathematics and vocabulary), will not change the point estimates much, because within the balanced groups, the distributions of the covariates are so similar—there are only small X differences for which to adjust. Nonetheless, it is still wise to apply those adjustments, because they typically do have minor, but beneficial, effects on point estimation (see, e.g., Rubin and Thomas 2000) and possibly substantial positive effects on estimated (not true) precisions of the point estimates, and thus the adjustments sharpen the resulting gold-standard interval estimates. These increases in estimated precisions arise because, in a general sense, estimated sampling errors are based on the estimated residual variances of outcomes conditionally given X within subclasses and strata, rather than on the unconditional variances of outcomes within subclasses and strata.

Of critical importance, all gold standard outcome analyses must be specified *before* any final outcome data are examined. Ideally, the model-based analyses will be prespecified in the protocol for the study, but some analyses may need to be specified later in the design phase, after some remaining imbalances in important covariates are identified. In any case, issues of dealing with possible multiple outcome analyses must be addressed, but these issues are beyond the scope of this discussion.

6. THE AFFECT OF PROPENSITY SCORE ANALYSES IN THE OBSERVATIONAL STUDIES ON GOLD STANDARD ANSWERS FROM RANDOMIZED EXPERIMENTS

The design and analysis of observational studies ideally follows the template of the design and analysis of randomized experiments, as I have strongly advocated recently (Rubin 2001, 2007, 2008). Here I simply address one issue in the analysis of observational studies that (curiously in some sense) affects the definition of gold standard answers from randomized experiments.

In observational studies, some types of units are commonly excluded from final outcome analyses because they have no approximately matching counterparts in the other treatment arm. Such cases may be revealed by the same sort of propensity score analyses as described in Section 4, and they definitely do occur with SCS's data set. For example, in SCS's observational study arm, nearly 25% of the subjects do not have overlapping estimated propensity scores (according to a particular sequential definition), due in part to the covariates "like math" and "prefer literature" with t statistics between the mathematics and vocabulary arms of well over ± 5. Of course, even when restricted to units with overlapping estimated propensity scores, such inferences for causal effects rely on the assumption of ignorable treatment assignment given X in the observational study arm of the experiment. This ignorability assumption seems plausible in SCS considering the extensive list of relevant X's that they collected, but entirely implausible given just their "covariates of convenience."

When we restrict the observational study to a subset of units (typically to those with overlapping estimated propensity scores), how are we to compare the observational study's thus-restricted answers to the unrestricted gold standard answers from the randomized experiment? In the article, both inferences are implicitly unrestricted in that, in both, SCS's analyses attempt to generalize to all units in their data base, no matter what their X values, even though the data suggest that all units with certain values of X would, with probability approaching 1, choose only one type of training class. Because in the article the

same set of X variables is available in the randomized experiment as in the observational study, and because these arms differ only randomly, the same estimated propensity scores can be calculated in the experiment as in the observational study. That is, all units in the experimental arm have estimated propensity scores for choosing mathematics versus vocabulary if given the choice, based on their X values using the propensity score function estimated from the units in the observational study arm. These estimated propensity scores in the randomized experiment allow us to restrict inferences in the randomized experiment to those types of units (defined by X) who would have clearly positive chances of preferring either mathematics or vocabulary training, thereby allowing valid comparisons of the inferences from the randomized and the observational study arms.

7. DISCUSSION

In this brief discussion, I have focused on the design and analysis of gold standard randomized trials. I have emphasized the importance of collecting extensive baseline covariates, both to make inferences more precise in the experiments and to allow analyses making comparisons between observational studies and randomized experiments to be more valid. I also have emphasized the critical role of design for balancing covariates across randomized treatment arms, both before final assignments and after final assignments. Once outcome data are revealed, I argue that the use of model-based adjustments (e.g., ANCOVA) is desirable, but more for obtaining improved estimates of precision, and thereby gold standard interval estimates, than for altering gold standard point estimates. Throughout, I have ignored complicating issues such as missing data, noncompliance, and dropout. Although, these are critically important issues that merit attention, they are beyond the scope of this short discussion.

The authors and the editorial board of *JASA* must be thanked for their wisdom in bringing this stimulating article to the attention of *JASA*'s readers.

ADDITIONAL REFERENCES

Ahmed, A., Husain, A., Love, T., Gambassi, G., Dell'Italia, L., Fracis, G., Gheorghiade, M., Allman, R., Meleth, S., and Bourge, R. (2006), "Heart Failure, Chronic Diuretic Use, and Increase in Mortality and Hospitalization: An Observational Study Using Propensity Score Methods," *European Heart Journal*, 27, 1431–1439.

Imbens, G., and Rubin, D. B. (2008), *Causal Inference in Statistics, and in the Social and Biomedical Sciences*, New York: Cambridge University Press.

Rubin, D. B. (2001), "Estimating the Causal Effects of Smoking," *Statistics in Medicine*, 20, 1395–1414.

——— (2007), "The Design versus the Analysis of Observational Studies for Causal Effects: Parallels With the Design of Randomized Trials," *Statistics in Medicine*, 26, 20–30.

——— (2008), "For Objective Causal Inference, Design Trumps Analysis," *The Annals of Applied Statistics*, 2, 808–840.

Rubin, D. B., and Thomas, S. N. (2000), "Combining Propensity Score Matching With Additional Adjustments for Prognostic Covariates," *Journal of the American Statistical Association*, 95, 573–585.

Wolos, C., and Rubin, D. B. (2008), "A Propensity Score Analysis to Improve Covariate Balance in a Randomized Experiment: Reducing Bias in the Prevention of Perinatal Sepsis (POPS) Trial," Poster JSM.

Zell, E. R., Kuwanda, M. L., Rubin, D. B., Cutland, C. L., Patel, R. M., Velaphi, S. C., Madhi, S. A., and Scharg, S. J. (2007), "Conducting and Analyzing a Single-Blind Clinical Trial in a Developing Country: Prevention of Perinatal Sepsis, Soweto, South Africa," *Proceedings of the International Statistical Institute, 56th Session* [CD–ROM].

Rejoinder

William R. SHADISH, M. H. CLARK, and Peter M. STEINER

We wish to thank Little, Long, and Lin (henceforth LLL), Hill, and Rubin for their thoughtful comments. It is gratifying that none of three commentaries had any fundamental concerns about our basic design. The three commentaries were largely orthogonal in focus, perhaps reflecting the divergent creativity that seems to characterize field experimentation today (Shadish and Cook 2009). Rubin suggests that we could have attended more closely to ensuring that the randomized experiment in our study was well balanced, as was the initial randomization to random or nonrandom assignment. These comments are sufficiently compelling that little rejoinder is needed. Time constraints on submitting this rejoinder prevent us from reporting the results of Rubin's recommended reanalysis of our data here, but we will do so in the foreseeable future. We make only one minor point here: Rubin stated that when he reanalyzed our data, "nearly 25% of the subjects do not have overlapping estimated propensity scores." Using our propensity scores, which differ from those estimated by Rubin, but using his sequential approach to discard nonoverlapping subjects, we find that only 9% of cases lack overlap. This illustrates that at least some of the decisions in a propensity score analysis are a matter of judgment to some extent, and much systematization of those decisions remains to be done. This requires continued research into the qualities that propensity score analyses need to be optimal.

But Rubin's recommendations for attending to balance in all of the randomizations involved in such studies are important not just for our study, but also for all researchers who compare results from nonrandomized and randomized experiments, whether the latter are called hybrid designs, doubly randomized preference designs, confirmatory evaluation studies, or single-study approaches. To the best of our knowledge, none of those

William R. Shadish is Professor, Psychological Sciences Section, School of Social Sciences, Humanities, and Arts, University of California, Merced, CA 95344 (E-mail: *wshadish@ucmerced.edu*). M. H. Clark is Assistant Professor of Psychology, Department of Psychology, Southern Illinois University, Carbondale, IL 62901 (E-mail: *mhclark@siu.edu*). Peter M. Steiner is Assistant Professor, Institute for Advanced Studies, 1060 Vienna, Austria, and currently a Visiting Research Scholar at Northwestern University, Evanston, IL 60208 (E-mail: *steiner@ihs.ac.at*). Shadish and Steiner were supported in part by grant 0620-520-W315 from the Institute for Educational Sciences, U.S. Department of Education.

Journal of the American Statistical Association
December 2008, Vol. 103, No. 484, Applications and Case Studies
DOI 10.1198/016214508000001020

studies has ever done the kind of statistical balancing that Rubin advocates. Their failure to do so calls into question whether the gold standard that has been used in all of these studies is 24 karat or some lesser degree of purity. Hill talks about the substantive importance of continuing to evaluate effect estimates from "less pristine" nonrandomized experiments. That may be true for some purposes, such as trying to provide the best possible answer to a substantive policy question given the available data. But if the goal is methodological, to determine whether there are quasi-experimental design and analysis practices that can well approximate results from randomized experiments, then tests in which either the randomized or the nonrandomized arms of the comparison are less than pristine are inherently limited in their capacity to inform the issue (Cook, Shadish, and Wong, 2008).

LLL reanalyzed our data with their approach to estimating both the effects of treatment and the effects of preference for treatment. The results are intuitive, and, if we understand correctly, provide an estimate of the effects of treatment on the treated, rather than the average causal effect that we computed. It is very helpful to see that both estimates can be computed from our design. Of as much interest to us, however, is LLL's comment that "studies with hybrid designs can answer scientific questions that otherwise may not be answered from completely randomized studies, and they are of potential interest to funding agencies." In our original discussion, we stated that we were unsure how fundable a hybrid design might be. Hill is even more pessimistic about funding in her comment. But LLL might rightly respond that estimating the effects of treatment on those who would choose that treatment is of direct policy relevance, because policymakers rarely create interventions that force people to participate. We may be able to get the same estimate from an experiment that randomizes those who prefer a treatment to that treatment or no treatment, because all participants then have the same preference. But few policy-relevant randomized experiments use a no treatment control; rather, they compare treatments to one another, where differential preferences may well exist, and some participants will be randomized to the condition they did not prefer. If so, preference effects might be best estimated with a hybrid design, and so those designs may be more fundable than we guessed.

LLL also note that "the study of Janevic et al. (2003) illustrates that the DRPT design can be successfully implemented and yield meaningful results regarding treatment preference in a real setting," in contrast to our method, which they describe as a "classroom exercise." Hill makes similar points. We think the choice of the term "meaningful" is correct; we also read the reanalyses of Janevic et al. as making perfect conceptual sense. But besides being meaningful, the results also must be accurate. In the face of nontrivial attrition and partial treatment implementation in these real-world settings, being certain of this accuracy is much harder. We tend to not question accuracy when results seem both meaningful and in the direction we expect, as with the study of Janevic et al. When results contradict our expectations, however, the dilemma is posed more pointedly, as we discuss in more detail at the end of this rejoinder.

Hill provides with a thoughtful overview of the diverse ways in which we can compare randomized and nonrandomized experiments. From it, we learned much about the kinds of comparisons of randomized and nonrandomized experiments that have been or can be done. Indeed, we suspect there are many more possibilities than either she or we have recognized, and it is good to see an effort to gather, categorize, and critique them. In the process of her overview, Hill makes four claims worthy of reply. We reply to these claims first, and then return to the more important general issue that both Hill and LLL raise about the relative value of pristine versus real-world comparisons of randomized and nonrandomized experiments.

First, Hill says that "Shadish, Clark, and Steiner (henceforth SCS) present their proposed CES as the ideal form in this genre." What we actually said was that our "method is just one alternative with its own strengths and weaknesses compared with previous methods." We hope no other readers would confuse Hill's statement with ours. Second, Hill says that "the SCS design (as proposed) is not useful for estimating such effects as the effect of the treatment on the treated, which often can be more useful from a practical or policy perspective." Before reading LLL's comment, we would have thought so as well. But this may be wrong if LLL's analysis provides a treatment on the treated (TOT) estimate. Moreover, our design could be adapted to provide the TOT estimator in the way in which Hill described. Earlier this year we submitted a grant proposal to conduct a large-scale study that is nearly identical to what Hill describes (although the grant proposed was rejected). Third, Hill asserts that we could have done more to explore the role of covariate choice in the accuracy of adjustments to nonrandomized experiments. We are currently doing such explorations (Steiner, Cook, Shadish, and Clark 2008). Two preliminary results are salient, although this is necessarily an oversimplified summary of a complex manuscript that is still being revised. One preliminary result is that obtaining good balance on covariates is necessary but far from sufficient for bias reduction, similar to the near-perfect balance but very poor bias reduction that we got using predictors of convenience. The other preliminary result is that bias reduction seems to be proportional to the extent to which the covariate set correlates with treatment choice and outcome, at least in our data.

Fourth, Hill dismisses the use of meta-analysis, saying that "combining results from 100 well-done studies may not help elucidate much of anything" regarding the extent to which nonrandomized experiments can approximate results from randomized experiments. We disagree. On the one hand, it is true that meta-analysis is mostly a correlational enterprise, with limits on the extent to which one can ever know for certain that one has correctly captured all of the study-level confounds with the assignment method. Indeed, that fact was the motivation for us to design the current study after years of studying the issue with meta-analytic methods (e.g., Heinsman and Shadish 1996; Kownacki and Shadish 1999; Shadish, Matt, Navarro, and Phillips 2000; Shadish and Ragsdale 1996). On the other hand, the correlational nature of the less pristine observational studies that Hill describes may not be much better. Moreover, in a recent review of both the published meta-analyses just cited and some related unpublished work in the first author's laboratory (Shadish 2008), we repeatedly found several findings that are meaningful in the sense that LLL used the term to describe their analysis of the work of Janevic et al. (2003). These include the following:

- Nonrandomized experiments that allow participants to self-select into conditions tend to yield less accurate results than those where nonrandomized selection is controlled by someone other than the participant.
- A standardized mean difference statistic calculated on pretest covariates related to outcome provides a plausible measure of the amount of selection bias in nonrandomized experiments.
- If nonrandomized experiments are conducted similar to randomized experiments in all aspects except assignment method, their results tend to be similar. Particularly important design features are using control groups from the same location and with the same substantive characteristics as the treatment group (focal local controls), demonstrating pretest equivalence on important covariates, and preventing self-selection into conditions.

In addition, both Kuss, Legler, and Borgermann (2008) and West and Thoemmes (2008) recently studied applying propensity score analyses to meta-analysis and found general similarities between results from randomized and nonrandomized experiments equated on covariates with propensity score stratification. Whether or not these results ultimately turn out to be accurate, they do cohere with our expectations about good design of nonrandomized experiments, as well as with similar results using other methods (e.g., Cook et al. 2008). Meta-analytic methods have a legitimate place in the body of confirmatory evaluation studies that Hill aspires to catalog.

But the crucial issue raised by Hill, and also by LLL, is the relative value of pristine versus real-world comparisons of randomized and nonrandomized designs. Less pristine studies currently constitute nearly all of the tests of accuracy of propensity score adjustments, and in many circles the consensus is that these tests suggest that propensity score adjustments do not work (e.g., Glazerman, Levy, and Myers 2003). Thus advocating less pristine studies for such tests is very risky, and is especially ill-advised if the studies are less pristine in ways that are confounded with the test of propensity scores. Speaking of one of her own studies in which propensity score adjustments seemed to mimic randomized results fairly well, Hill states that "we cannot know whether convergence of estimates in scenarios such as this necessarily implies that the observational study was sufficient to estimate causal effects." What is more important, however, is that the opposite also is true. We cannot know whether lack of convergence of estimates in such scenarios necessarily implies that the observational study was insufficient to estimate causal effects. In comparisons where either the randomized or the nonrandomized arms are less than pristine, it is too easy to find potential alternative explanations for the lack of convergence.

This is well illustrated by the study of Peikes, Moreno, and Orzol (2008) that Hill also cites at the end of her comment. Ironically, in the present context, the study of Peikes et al. does claim to have tested propensity score matching under "ideal conditions" (p. 222), finding that propensity score matching yielded incorrect results compared with randomized experiments. The study of Peikes et al. (2008) was admirable in many respects and is a welcome addition to this literature, but it is hardly ideal. For example, their nonrandomized comparison group was from a different location than the randomized group and so must be presumed to be from a different population subject to different unmeasured influences. Ideally, the comparison group is from the same place and population as the randomized experiment, differing only in not being randomized. Peikes et al. noted that their covariates may not have included key predictors of why people chose to enter the program; ideally, they would taken such actions as interviewing participants to find out what influenced them to enter the study and the intervention, and then measuring those factors, to improve the plausibility of the strong ignorability assumption. They used older methods to test balance (Rosenbaum and Rubin 1984) and had a few imbalances on key covariates; ideally, they would have used newer balancing procedures (e.g., Rubin 2001) and removed all imbalances on all key covariates. One of their three sites had a very small sample size—$N = 22$ in treatment and $N = 19$ in the propensity score–matched comparison group—that no propensity score analyst would see as ideal. They acknowledged the small sample size but still presented the site results as evidence that propensity score analyses do not work. Some of these flaws are remediable, but not all, so that the study of Peikes et al. can never provide a test of propensity score matching under ideal conditions.

Compared with our study, then, the intellectual dilemma for Hill and LLL posed by the studies of Peikes et al. (2003) and Arceneaux, Gerber, and Green (2006) is that they are exactly the kind of less pristine observational studies that Hill and LLL suggest might be meaningful. If they are, does this mean that propensity score adjustments do not work (at least in those single instances)? We think not, because propensity score adjustments were not designed to adjust for the less pristine aspects that have little to do with selection mechanisms. Rather, they aim to estimate what would have happened to a set of nonrandomized participants had they been randomized with all other features of the study remaining the same, not what would have happened had some other participants from a different population in a different place been randomized followed by attrition that might be differential. If this statement is true, then the lessons from less pristine studies will always be more ambiguous that the results from designs like ours.

Of course, we do not expect our design to be widely used in real-world evaluations. Indeed, using it in that fashion might well destroy the very pristineness that is its advantage, by introducing significant problems of attrition, missing data, and the like. In that sense, we view the design as a cross between a computer simulation and a field experiment. Just as we would not recommend that most investigators evaluate their real-world intervention with a computer simulation, we expect that they will rarely evaluate it with our design. Rather, it is just one more useful tool in our repertoire for methodologists who want to study the conditions under which results from nonrandomized experiments can approximate results from randomized experiments.

ADDITIONAL REFERENCES

Cook, T. D., Shadish, W. R., and Wong, V. C. (2008), "Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings From Within-Study Comparisons," *Journal of Policy Analysis and Management*, 27, 724–750.

Kownacki, R. J., and Shadish, W. R. (1999), "Does Alcoholics Anonymous Work? The Results From a Meta-Analysis of Controlled Experiments," *Substance Use and Misuse*, 34, 1897–1916.

Kuss, O., Legler, T., and Borgermann, J. (2008), "Do Randomized and Non-Randomized Trials Yield Different Answers in Similar Populations? Evidence From a "Meta-Propensity Score" Analysis in Cardiac Surgery," presented at the Fourth Symposium on Causality, Jena, Germany, available at *http://www.metheval.uni-jena.de/projekte/symposium2008/program.php?&lang=en*.

Peikes, D. N., Moreno, L., and Orzol, S. M. (2008), "Propensity Score Matching: A Note of Caution for Evaluators of Social Programs," *The American Statistician*, 62, 222–231.

Pohl, S., Eisermann, J., and Soellner, R. (2008), "A Randomized Experiment Comparing Random to Nonrandom Assignment: A Replication of Shadish and Clark," presented at the Fourth Symposium on Causality, Jena, Germany, available at *http://www.metheval.uni-jena.de/projekte/symposium2008/program.php?&lang=en*.

Shadish, W. R. (2008), "An Empirical Program of Quasi-Experimentation," presented at the Fourth Symposium on Causality, Jena, Germany, available at *http://www.metheval.uni-jena.de/projekte/symposium2008/program.php?&lang=en*.

Shadish, W. R., and Cook, T. D. (2009), "The Renaissance of Field Experimentation for Evaluating Interventions," *Annual Review of Psychology*, 60.

Shadish, W. R., and Ragsdale, K. (1996), "Random Versus Nonrandom Assignment in Controlled Experiments: Do You Get the Same Answer?" *Journal of Consulting and Clinical Psychology*, 64, 1290–1305.

Shadish, W. R., Matt, G. E., Navarro, A. M., and Phillips, G. (2000), "The Effects of Psychological Therapies Under Clinically Representative Conditions: A Meta-Analysis," *Psychological Bulletin*, 126, 512–529.

Steiner, P. M., Cook, T. D., Shadish, W. R., and Clark, M. H. (2008), "Differential Role of Covariate Selection in Controlling for Selection Bias in Observational Studies," unpublished manuscript, Institution for Policy Research, Northwestern University.

West, S. G., and Thoemmes, F. (2008), "Observational Studies: Towards Improving Design and Analysis," presented at the Fourth Symposium on Causality, Jena, Germany, available at *http://www.metheval.uni-jena.de/projekte/symposium2008/program.php?&lang=en*.