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

[illegible]

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

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,

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 bothe.g., dozon

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 rom

At the end of the time allotted for pretests, participants were assigned randomly to be in a randomized (= 235) or a nonrandomized (= 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 (= 116). Participants born in three other randomly chosen months were sent to the mathematics training condition of the randomized experiment (= 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.

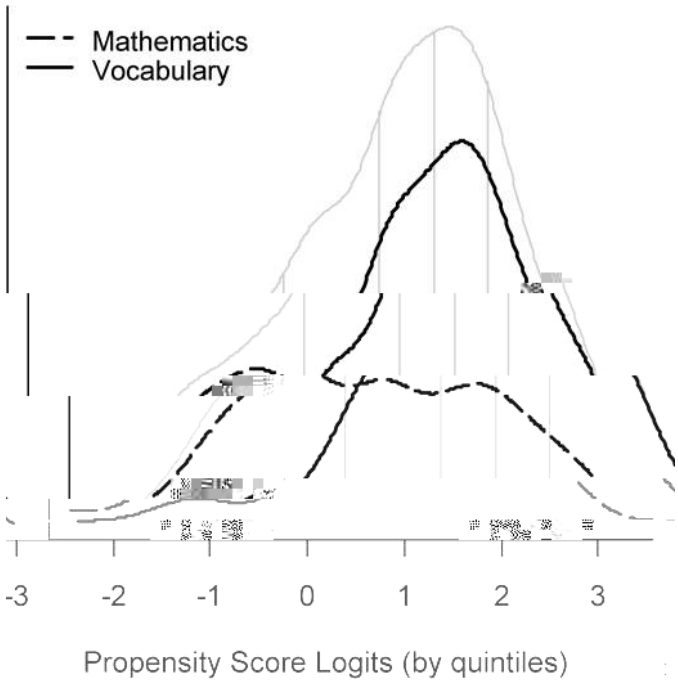
Results from the randomized experiment are the presumed best estimate against which all adjusted and unadjusted nonrandomized results are compared. But randomized experiments still encounter group differences in covariates due to sampling error, so we adjusted the randomized results using all of the 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


Mathematics
Vocabulary
-3
-2
-1
0
1
2
3
Propensity Score Logits (by quintiles)

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.]

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*