Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random to Nonrandom Assignment

William R. Shadish

University of California, Merced

M.H. Clark

Southern Illinois University, Carbondale

Peter M. Steiner

Institute for Advanced Studies, Vienna, Austria

To appear in the *Journal of the American Statistical Association*

Authors' Footnote

William R. Shadish is Professor and Founding Faculty, University of California, PO Box 2039, Merced CA 95344 (wshadish@ucmerced.edu); M.H. Clark is Assistant Professor of Psychology, Mailcode 6502, Southern Illinois University, Carbondale IL 62901 (mhclark@siu.edu); and Peter M. Steiner is Assistant Professor at the Institute for Advanced Studies, Stumpergasse 56, 1060 Vienna, Austria, and currently a visiting research scholar at Northwestern University, Evanston, IL 60208 (steiner@ihs.ac.at). The first and third authors were supported in part by grant R305U070003 from the Institute for Educational Sciences, U.S. Department of Education.

Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random to Nonrandom Assignment

Abstract

A key justification for the use of 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, past 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 or a nonrandomized experiment. In the randomized experiment, participants were randomly assigned to mathematics or vocabulary training; in the nonrandomized experiment, they 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 outcome. Ordinary linear regression reduced bias in the nonrandomized experiment 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 scores; Randomized Experiment; Selection bias.

Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random to Nonrandom Assignment

1. INTRODUCTION

Randomized experiments can yield unbiased estimates of effect sizes. However, 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, prior 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 we will 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 impact 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. However, 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 to 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 to the nonrandomized control on the assumption that if the two control groups are equal, the nonrandomized control could be substituted for the randomized control). This method gives the researcher access to raw data from individual participants, so the researcher 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. For example, Heckman, Ichimura and Todd (1997) randomly assigned applicants to a control group or to a job training program. In addition, data were collected on a group of eligible nonparticipants who met the requirements for the training program but were not participating in it. Heckman et al. 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. By comparison, more optimistic results were obtained in studies by Dehejia and Wahba (1999) and by 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 occurs both pre-treatment and post-treatment in field experiments.

At first glance, studies like 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 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 trying to estimate *what the effect would have been if the participants in a nonrandomized experiment had instead been randomly assigned to the same conditions using the same measures at the same time and place*. The latter counterfactual cannot be directly observed. 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, where all other features of the experiment are held equal. This was not done in Dehejia and Wahba (1999), Heckman et al. (1997) and Hill et al. (2004), or any other such studies. Rather, assignment mechanism (random or nonrandom) varied nonrandomly in those studies, and is always confounded with other differences between the random and nonrandom control groups. For example, compared to the randomized control group, the nonrandomized control group is often 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, we cannot know if 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) use the simplest form of this approach, summarizing results from dozens of meta-analyses comparing randomized and nonrandomized experiments. The average over these comparisons was zero—nonrandomized experiments yielded the same effect size as randomized experiments on average—though in any given meta-analysis the difference usually was not zero. However, 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 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, some concluding that the adjusted difference is near zero (Heinsman and Shadish, 1996) and others concluding 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 properly modeled. Moreover, the meta-analytic approach also cannot access primary raw data from each experiment, so it cannot test whether adjustments such as selection bias modeling or propensity score analysis improve estimates from nonrandomized experiments.

To address some of the problems with these past methods, the present study explores the differences between randomized and nonrandomized experiments using a laboratory analogue that randomly assigns participants to be in either randomized or nonrandomized experiments that are otherwise 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 et al, 1993; Janevic et al., 2003; Long, Little and Lin, in press)—though it differs in the following 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 below, usually impede the ability of DRPTs to obtain a good test of the effects of adjustments like propensity score analysis. Second, none of the DRPT studies done 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 if participants had been randomized.

Third, and perhaps most important, because the present method uses a brief laboratory analogue treatment, it avoids problems of partial treatment implementation and of missing outcome data that have occurred in the few past DRPTs that have actually tried to gather data. This is crucial because adjustments like propensity score analysis only answer questions 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 the latter problems. For example, nearly two-thirds of those initially assigned to conditions in Janevic et al. (2003) refused to accept their random assignment to the randomized or choice arms of the study, and all of these two-thirds had missing outcome data. Although the differential rate of refusal (3%) to conditions is minimal (62% refusal to the choice arm versus 65% to the randomization arm), an additional 4% withdrew from the choice arm after pretest, making 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 report 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. (in press) reports using an intent to treat analysis in order 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. So the resulting comparison of the adjusted results from the Janevic et al. (2003) randomized and nonrandomized experiments is a joint test of the effects of adjusting for assignment method, missing outcomes, and partial treatment implementation. The present method substantially avoids the latter two problems, and thus allows for tests of the effects of adjustments for assignment method that are less encumbered by extraneous concerns.

The present 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 we describe more in Section 4. In addition, the present 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. Hence the present method is just one alternative with its own strengths and weaknesses compared to past methods.

Still, the unique contribution of this study is its 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 to be interested in a brief laboratory analogue treatment, this format is a key virtue because it allows estimates 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 experiment. Though 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 in order to explore its generalizability.

2. METHODS

The study began with baseline tests that were later used to predict treatment selection (see Figure 1). Then participants were randomly assigned to be in a randomized or 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 for assignment method.

Insert Figure 1 about here

2.1. Participants

Volunteer undergraduate students from introductory psychology classes at a large Midsouthern 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, and described in more detail below. These sample sizes are not large, a limitation if propensity scores are most effective with large samples. However, 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 who signed up for the experiment, 445 completed 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, one to vocabulary), and two were randomly assigned to the non randomized experiment (one chose vocabulary, one not complete the choice form). These five 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.

2.2. Pretests

Written instructions and computer scored answer sheets were used for all of the following pretests: (1) Demographics Questionnaire I, prepared by the present researchers, gathered data about participant age, education, marital status, major area of study, ACT and

SAT scores, GPA for college and high school; (2) The Vocabulary Test II (Educational Testing Services, 1962) measured vocabulary skills to predict selection into mathematics or vocabulary training; (3) The Arithmetic Aptitude Test (Educational Testing Services, 1993), administered with scratch paper, measured mathematics skills to predict selection into conditions; (4) Demographics Questionnaire II, prepared by the researchers based on an interview with a full-time staff member of the student educational advising center, assessed prior scholastic experiences in mathematics and vocabulary to predict selection into condition; (5) The International Personality Item Pool test (Goldberg, 1997) assessed five major domains of personality: extroversion, emotional stability, agreeableness, openness to experience, and conscientiousness; (6) The Short Mathematics Anxiety Rating Scale (Faust, Ashcraft and Fleck, 1996) assessed stress induced by mathematics to predict selection into mathematics training; and (7) 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 five 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 analogues 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 prior to 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 and so making it easier to detect differences between randomized versus self-selected conditions.

Training sessions were conducted by one of four white males, three of whom were psychology graduate students and the other an 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 presented in training and 15 new) and 20 mathematics items (10 presented 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. However, we later found that the correct response for two mathematics items was not listed, so those items were removed from analyses.

2.5. Procedure

Data collection spanned 22 weeks, with 24 testing sessions having from 7 to 48 people per session. Participants signed up for the experiment between four weeks to one hour prior to participating. Upon 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 grade point averages (GPAs), college GPAs, mathematics and English grades, and ACT or SAT college admission scores; 92% of the participants consented. However, university records reported ACT scores for only 61.5% of 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: $N = 130$ (62%) of the quasi-experimental participants had complete predictor data, $N = 24$ (11%) had missing data on one predictor, and $N = 56$ (27%) had missing data on more than one predictor. However, the overall number

of missing observations was quite low (2.6% and 3.6% of all covariate measurements of the randomized and quasi-experiment, respectively). Therefore, to maintain the focus on the simple comparison of randomized and non-randomized 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, N = 131 (62.4%) chose vocabulary and N = 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. Afterwards, 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. However, randomized experiments still encounter group-differences in covariates due to sampling error, so we adjusted the randomized results using all the available covariates in backward stepwise regression. However, eventual bias reductions were similar whether we used the adjusted or unadjusted randomized results as a benchmark.

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 did participants who received vocabulary training (see Table 1). In the 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 ($\Delta = \left| 4.01 - 5.01 \right| = 1.00$ is a measure of the bias in the unadjusted nonrandomized results, where $\Delta = 0$ would indicate no bias.

Insert Table 1 about here

3.1.2. The Effects of Vocabulary Training on Vocabulary Outcome. In the covariance-adjusted randomized experiment, participants who received vocabulary training performed 8.25 points (out of 30) better on the vocabulary outcome than did 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 $\Delta = \left| 8.25 - 9.00 \right| = .75$.

3.2. Adjusted Results

There is only borderline evidence that the results from the nonrandomized experiment are 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 extend 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 second best for vocabulary outcome.

3.2.2. Using Propensity Scores. Though several other kinds of adjustments are possible, such as econometric selection bias modeling (e.g., Heckman, Ichimura and Todd, 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, …, N) let Z_i denote the treatment assignment ($Z_i = 1$ if the person receives treatment, in our study vocabulary training, $Z_i = 0$ if the person receives no or another treatment, here mathematics training) and $\mathbf{x}_i$ 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 is assumed that, given the $\mathbf{X}$'s, the Z_i are independent. Various authors (e.g., Rosenbaum & 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 & 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 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 zero, (b) the ratio of the variance of the propensity score in the two groups (R) should be near one, and (c) ratio of the variances of the covariates after adjusting for the propensity score

must be close to one, where ratios between 0.80 and 1.25 are desirable, and those smaller than 0.50 or greater than 2.0 are far too extreme. The propensity scores 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 were also well-balanced using the criteria proposed by Rosenbaum and Rubin (1984)—a 2 x 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 5 units in each cell.

Insert Tables 2, 3, and Figure 2 about here

Table 1 reports four propensity score adjustments for the nonrandomized experiment: (a) stratification on propensity score quintiles (Rosenbaum & 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 lines reading “Plus Covariates”). The table also reports the usual regression-based standard errors, except that standard errors for methods involving

propensity scores were bootstrapped (for each bootstrap sample the propensity scores were re-fit; predictors included remained unchanged).

Overall, the eight propensity score adjustments reduced bias by an average of 74% (range 59-96%), depending on the model. Bias reduction was higher for vocabulary (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 increased the variance accounted for substantially, made little difference to bias reduction, and reduced the bootstrapped standard errors of the estimate slightly. Standard errors for propensity score weighting were larger than for any other method, probably inflated by the presence of some very small propensity scores. Standard errors were also 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, standard errors for propensity score adjusted effects were moderately larger than for the original covariate-adjusted randomized experiments.

Selection of covariates to use in creating propensity scores is a crucial feature of good propensity score analysis (Brookhart, Schneeweiss, Rothman, Glynn, Avor and Stürmer, 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 sex, age, marital status and race (dummy coded for two predictors: Caucasian, and

African American) as predictors. Those variables are often 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); and 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 are clearly 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 line of data). However, this would have hidden a failure to balance very well on many of the remaining covariates that would presumably have been unobserved by such a researcher (Table 3, fourth line of data). 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, but also 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 (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 do even better if the study were 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 shown by our analysis of propensity scores based on predictors of convenience, 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 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 Rosenbaum and Rubin (1984) criteria on significance testing makes it vulnerable to confusing successful balance with low power. The emphasis in Rubin (2001) on the size of imbalance may be more desirable; and both sets of criteria should probably be reported. We would benefit from further development of ways to create and assess balance (e.g., Imai, King & Stuart, 2007; Sekhon, 2007), and from better-justified standards for how much balance should be achieved.

Results were also sensitive to how missing data in the predictors were managed. At first, we followed one of Rosenbaum and Rubin's (1984) recommendation to create propensity scores separately for groups with different missing data patterns. However, we found that bias reduction was highly sensitive to seemingly minor changes in how those patterns were identified, in one case even increasing bias. Hence we moved to more current missing data methods, but those results may also 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, such as 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 may also be sensitive to which of these methods is 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 be yielding adjustments of unknown or highly variable accuracy. For a method as new as propensity score analysis, this is not surprising, 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 did as well or better. One situation for using propensity scores 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.

4.2. Comments on the Laboratory Analogue Design Used in This Study

Questions may arise about the replicability and generalizability of these results given the design used. The design is probably 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. So 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 added achievement motivation as an additional predictor of selection. The first author is working to computerize administration of this method, which might allow more complex assignment mechanisms to be quickly implemented, 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 to do random assignment rather than using birth month.

The question of generalization is more serious, and has two parts. The first part concerns how the results reported in this study would change over variations of the method that stay within this general laboratory analogue paradigm. One could vary the kind of

treatment from the current educational one to mimic other substantive areas such as job training, health, or different parts of education. Similarly, one could create more time-consuming treatments, although it would be desirable to avoid attrition from both treatment and from measurement because they are separate problems from adjusting for selection into conditions.

A second variation within the laboratory analogue method is to study different selection mechanisms, such as cutoff-based assignment mechanisms used in the regression discontinuity design (Shadish et al., 2002), analogues to parental selection of children into interventions, or analogues 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 within 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 are also 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 would also decrease the 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 within this method is to study people 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 one-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 analogues 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 replicate 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 analogue 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. However, 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 Janevic et al., 2003 study (see also Long, Little and Lin, in press). 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 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.

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 technic. *Postgraduate Medicine, December, 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 Corporation.

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 non-experimental studies: Re-evaluating 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*(2), 24-33.

Educational Testing Service. (1962). Vocabulary Test II (V-2). *Kit of factor referenced cognitive tests*. Princeton, NJ: Educational Testing Services.

Educational Testing Service. (1993). Arithmetic Aptitude Test (RG-1). *Kit of factor referenced cognitive tests*. Princeton, NJ: Educational Testing Services.

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* [on line]. Available: http://ipip.ori.org/ipip/appendixa.htm#AppendixA.

Greene, W.H. (1999). *Econometric Analysis*. Upper Saddle River, New Jersey: 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 A. Gelman and X-L. Meng (Eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete Data Perspectives* (pp. 51-60). New York: John Wiley & Sons.

Hill, J.L., Rubin, D.B., and Thomas, N. (2000). The design of the New York School Choice Scholarship program evaluation. In L. Bickman (Ed.), *Validity and Social Experimentation: Donald Campbell's Legacy* (Vol. 1) (pp. 155-180). Thousand Oaks, California: Sage Publications.

Hirano, K., & 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., & Stuart, E. A. (2007). *Misunderstandings among experimentalists and observationalists about causal inference*. Retrieved July 20, 2007, from http://imai.princeton.edu/research/files/matchse.pdf.

Janeic, 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*(7), 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. (in press). Causal inference in hybrid intervention trials involving treatment choice. *Journal of the American Statistical Association*.

Luellen, J.K. (2007). *A Comparison of Propensity Score Estimation and Adjustment Methods on Simulated Data*. Unpublished doctoral dissertation, The University of Memphis, Memphis, Tennessee.

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., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70* (1), 41-55.

Rosenbaum, P.R., and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516-524.

Rosenbaum, P.R., and Rubin, D. B. (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.

Rubin, D.B. (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*. Retrieved July 20, 2007, from http://sekhon.berkeley.edu/papers/SekhonBalanceMetrics.pdf.

Shadish, W.R. (2000). The empirical program of quasi-experimentation. In L. Bickman (Ed.), *Validity and Social Experimentation: Donald Campbell's Legacy* (pp. 13-35). Thousand Oaks, California: Sage Publications.

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

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.

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*(4). AS56-AS66.

Wennberg, J.E., Barry, M.J., Fowler, F.J., & Mulley. A. (1993). Outcomes research, PORTs, and health care reform. *Annals of the New York Academy of Sciences, 703*, 52-62.

Table 1. Percent Bias Reduction in Quasi-Experimental Results by Propensity Score (PS) Adjustments

Mathematics Outcome

	Mean Difference (standard error)	Absolute Bias (Δ)	Percent Bias Reduction (PBR)	R^2
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%	.29
Plus Covariates	3.74 (.42)	.27	73%	.66
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%[a]	.35
ANCOVA Using Observed Covariates	3.85 (.44)	.16	84%	.63

Vocabulary Outcome

	Mean Difference (standard error)	Absolute Bias (Δ)	Percent Bias Reduction	R^2
Covariate-Adjusted Randomized Experiment	8.25 (.37)			.71
Unadjusted Quasi-Experiment	9.00 (.51)	.75		.60
PS Stratification	8.15 (.62)	.11	86%	.55
Plus Covariates	8.11 (.52)	.15	80%	.76
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. For propensity score stratification stratum weights according to propensity score quintiles were used. Standard errors for propensity score methods are based on 1,000 bootstrap samples (separate samples for each group), with re-fitted 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).

[a] This adjustment increased bias by 5%.

Table 2. Correlations between Predictors and Outcome in Nonrandomized Experiment

Predictor	Vocabulary Posttest	Mathematics Posttest	Chose Vocabulary Training
Vocabulary Pretest	.468**	.109	.169*
Mathematics Pretest	.147*	.446**	-.090
Number of Prior Mathematics Courses[a]	-.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[a]	.087	.149*	-.014
Caucasian	.322**	-.074	.178*
African-American	-.296**	-.015	-.144*
Age[a]	.077	-.217**	.022
Male	.064	.141*	-.065
Married	-.073	-.162*	.001
Mother Education	.094	-.022	.010
Father Education	.110	.068	.008
College Credit Hours[a]	.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)

[a] These four variables were log-transformed in all analyses to reduce positive skew.

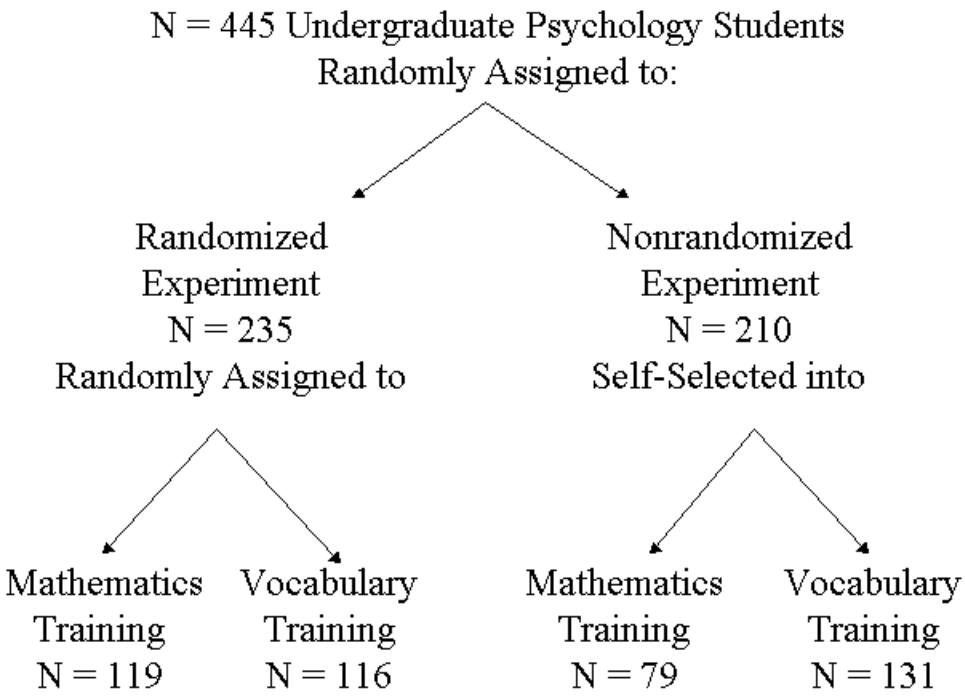
Table 3. Rubin's (2001) Balance Criteria Before and After Propensity Score Stratification

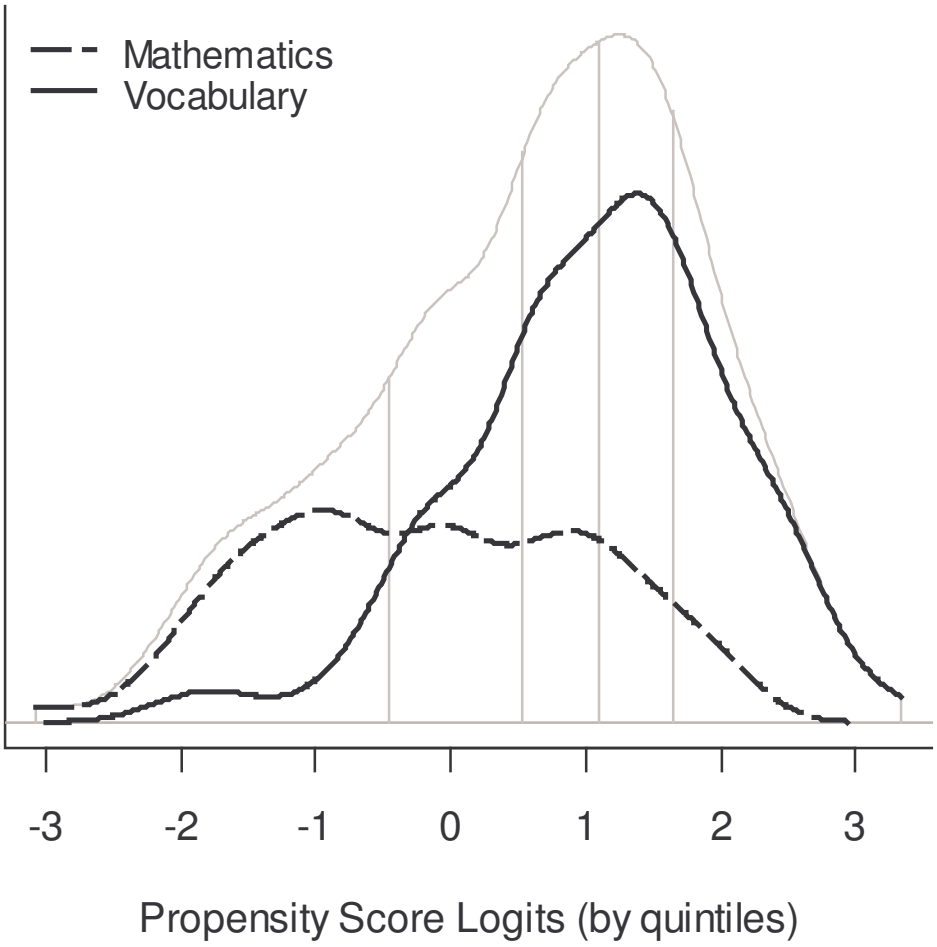
Propensity Score			Number of Covariates with Variance Ratio				
Analysis	B	R	≤1/2	>1/2 and ≤4/5	>4/5 and ≤5/4	>5/4 and ≤2	>2
Before Any Adjustment							
	-1.13	1.51	0	2	17	6	0
After Stratification on Propensity Scores Constructed from All Covariates							
	-0.03	0.93	0	1	22	2	0
After Stratification on Propensity Scores Constructed from Predictors of Convenience							
Balance Tested only on the 5 Predictors of Convenience							
	-0.01	1.10	0	0	5	0	0
After Stratification on Propensity Scores Constructed from Predictors of Convenience							
Balance Tested on All 25 Covariates							
	-0.01	1.10	0	2	16	7	0

Note. Standardized mean difference in propensity scores are 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 1. Overall design of this study.

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


N = 445 Undergraduate Psychology Students
Randomly Assigned to:
Randomized Experiment
N = 235
Randomly Assigned to
Nonrandomized Experiment
N = 210
Self-Selected into
Mathematics Training
N = 119
Vocabulary Training
N = 116
Mathematics Training
N = 79
Vocabulary Training
N = 131


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