Published in final edited form as:

Psychol Methods. 2012 June ; 17(2): 244–254. doi:10.1037/a0028031.

Estimating the Causal Effect of Randomization Versus Treatment Preference in a Doubly Randomized Preference Trial

Sue M. Marcus,

Departments of Psychiatry and Biostatistics, Columbia University, and New York State Psychiatric Institute, New York, New York

Elizabeth A. Stuart,

Johns Hopkins Bloomberg School of Public Health

Pei Wang,

Department of Biostatistics, Columbia University

William R. Shadish, and

Department of Psychological Sciences, University of California at Merced

Peter M. Steiner

Department of Educational Psychology, University of Wisconsin–Madison

Abstract

Although randomized studies have high internal validity, generalizability of the estimated causal effect from randomized clinical trials to real-world clinical or educational practice may be limited. We consider the implication of randomized assignment to treatment, as compared with choice of preferred treatment as it occurs in real-world conditions. Compliance, engagement, or motivation may be better with a preferred treatment, and this can complicate the generalizability of results from randomized trials. The doubly randomized preference trial (DRPT) is a hybrid randomized and nonrandomized design that allows for estimation of the causal effect of randomization versus treatment preference. In the DRPT, individuals are first randomized to either randomized assignment or choice assignment. Those in the randomized assignment group are then randomized to treatment or control, and those in the choice group receive their preference of treatment versus control. Using the potential outcomes framework, we apply the algebra of conditional independence to show how the DRPT can be used to derive an unbiased estimate of the causal effect of randomization versus preference for each of the treatment and comparison conditions. Also, we show how these results can be implemented using full matching on the propensity score. The methodology is illustrated with a DRPT of introductory psychology students who were randomized to randomized assignment or preference of mathematics versus vocabulary training. We found a small to moderate benefit of preference versus randomization with respect to the mathematics outcome for those who received mathematics training.

Keywords

generalizability; causal inference; conditional independence; propensity score matching; treatment preference

Since the early days of statistics, there has been continued interest in whether treatments studied under rigorous scientific conditions yield causal estimates that generalize to real-world scenarios (Fisher, 1935). Evidence from a scientifically rigorous trial is relevant only to the extent that it facilitates generalization or extrapolation from the estimated experimental causal effect to a real-world setting (e.g., Flay, 2005; Levitt & List, 2007; Shadish, Cook, & Campbell, 2002). Unless there are formal approaches for generalizing from rigorous trials to real-world conditions, there may be inappropriate, arbitrary, or uneven use of treatments in practice (Braslow et al., 2005).

In particular, one factor that may lead to variation in causal effects is the randomization itself: Individuals who select their treatment may have different outcomes as compared with individuals randomly assigned to that same treatment (Marcus & Gibbons, 2002; Shadish et al., 2002). In this current article, our goal is to address whether causal estimates derived from randomized trials generalize to the more real-world scenario in which treatment is assigned by preference.

Differences in outcomes for those who are randomized to treatment versus those who choose their treatment may be due, in part, to selection factors; that is, those who choose a particular treatment may tend to have specific characteristics. Beyond selection effects, psychological factors such as motivation, engagement, and compliance may be better with a preferred rather than randomized treatment, compromising the generalizability of results from randomized trials (Macias et al., 2009). The randomization itself may impact a person's psychological or social response to a treatment (Shadish et al., 2002). For example, the Women Take Pride study assessed group versus self-directed behavioral interventions for women with heart disease (Janevic et al., 2003) and found much higher adherence rates for the preferred interventions (Long, Little, & Lin, 2008). If the mode of assignment, randomization or self-selection, affects the outcome via such psychological factors, it can also be seen as a violation of the stable-unit-treatment-value assumption (SUTVA), which is a fundamental assumption underlying causal inference. SUTVA basically states that the potential outcomes do not depend on the assignment mechanism and other subjects' assignment (i.e., not interference between units).

Heterogeneity in outcomes might also be due to heterogeneity in the treatment effect (see, e.g., Heckman, 2001). The idea that the same intervention may have different outcomes, even after adjusting for a set of population characteristics, has received much attention in the econometrics literature. In the "model of essential heterogeneity," responses to treatment are assumed to be heterogeneous, treatment choices are based in part on this heterogeneity, and some components of heterogeneity are unobserved. Thus, effectiveness of an intervention may vary, depending on a variety of factors that may be, at least in part, unobserved. For instance, if subjects choose the treatment based on the treatment effect they expect for themselves (which is, in general, not observable), those in a preferred treatment condition will very likely exhibit larger treatment effects on average than those in a randomized treatment condition.

In many studies, psychological factors such as motivation may differ across randomization versus preference settings, but may be unmeasured or poorly measured. A first step might be to see whether causal effects differ across these two settings without specifically attributing this to observed factors. This type of comparison provides evidence of experimental versus real-world differences in outcomes, but does not require explicit measurement of the psychological factors such as motivation, persistence, and engagement that may contribute to the differences.

Hybrid Randomized and Nonrandomized Designs

We introduce this section with a review of various hybrid randomized and nonrandomized designs that have been used to address generalizability. We end this section with a description of the doubly randomized preference trial (DRPT), a hybrid randomized and nonrandomized design that allows for the unbiased estimation of the causal effect of randomization versus treatment preference, the goal that was motivated above.

Although Sir Ronald A. Fisher (1935) was a proponent of the randomized experiment, he also criticized its use, in that the clear causal inference from the randomized experiment can sometimes come at the expense of generalizability of the estimated causal effect derived from the randomized experiment. Observational studies (in which the preferred treatment is chosen) may be better at estimating effects that generalize to real-world settings; however those effects are more subject to selection bias (Imai, King, & Stuart, 2008).

Various hybrid randomized and nonrandomized designs have been developed to “get the best of both worlds,” taking advantage of the increased internal validity from randomized assignment and the increased external validity from observational studies. These designs are variations on Solomon’s four-group design created by substituting different design elements for the pretest (Solomon, 1949).

Strategies for assessing generalizability of randomized controlled trials (RCTs) have focused on assessing whether RCT populations differ from target populations of real-world interest. Stevens et al. (2007) compared characteristics of children from the randomized Multimodal Treatment Study for Children with ADHD to those from the more representative National Institute of Mental Health Methods for the Epidemiology of Child and Adolescent Mental Disorders. Greenhouse, Kaizar, Kelleher, Seltman, and Gardner (2008) illustrated their approach for making generalizability judgments using a case study of the risk of suicidality among pediatric antidepressant users. Although these methods are useful in identifying conditions under which generalizability seems plausible, they do not propose strategies to deal with situations in which generalizability does not hold. Marcus (1997) provided methods to formally test for generalizability bias and also to derive an unbiased estimate of effectiveness for a target population. In the current article, we extend this approach to assess the causal effect of randomization versus choosing a preferred treatment.

Zelen (1990) proposed a randomized consent design that first randomizes subjects to the treatment and control conditions. Subjects randomized to treatment are then asked to give consent to receive the treatment. The subject is given the treatment if the subject gives consent and is given the control otherwise. The procedure is then followed similarly for the control condition. This design can be more powerful than a traditional randomized design when the traditional design is restricted to subjects who consent to randomization, but there is a question about whether it is ethical to assign the treatments before describing them (Ellenberg, Finkelstein, & Schoenfeld, 1992).

Another hybrid design is the parallel randomized and nonrandomized trial (Marcus, 1997; Paradise et al., 1984), also called the partially randomized preference trial (PRPT; Brewin & Bradley, 1989; Long et al., 2008). Generally, most trials exclude subjects who do not give consent for randomization, which is another factor that reduces the generalizability of results from randomized trials. In the PRPT, subjects who give consent for randomization are randomized to treatment versus control conditions. Those who do not give consent for randomization are instructed to choose treatment or control conditions and are followed similarly to those in the randomized portion. For example, in the Coronary Artery Surgery Study (1984), two thirds of the 2,099 subjects who met eligibility criteria refused to be randomized, so generalizability was uncertain. The use of the PRPT allowed investigation of

this issue by allowing nonexperimental estimation of effects for the group of subjects who did not consent to randomization. Nonconsent bias was also an issue for the study of surgery versus medication for otitis media (recurring ear infections), as the parents of the less severely affected children tended to prefer medication over surgery (Marcus, 1997; Paradise et al., 1984, 1990). Generally, most trials exclude subjects who do not give consent for randomization. However, when the randomized and nonrandomized data can be combined, the PRPT can increase accrual and power and is useful for assessing nonconsent bias or the generalizability of results to individuals who did not consent to randomization (Marcus, 1997).

The current article concerns another type of hybrid randomized and nonrandomized design called the DRPT (Long et al., 2008; Rucker, 1989; Shadish, Clark, & Steiner, 2008). In the DRPT, subjects are randomized either to randomization to treatment or to control or preference of treatment or control. The DRPT allows for estimation of preference effects and causal effects in subclasses defined by preference—effects that cannot be estimated from a randomized trial alone (Long et al., 2008). Furthermore, we show in this article that the DRPT allows for the unbiased estimation of the causal effect of randomization versus preference of treatment versus control.

We illustrate the methodology using a DRPT of introductory psychology students who were randomized to randomization or preference for either math training ($z = 0$) or vocabulary training ($z = 1$) (Shadish et al., 2008). Interestingly, this study has already been used to answer two questions (Long et al., 2008; Shadish et al., 2008). In this current article, we ask a third question. Shadish et al. (2008) asked whether nonrandomized studies can yield answers that are similar to randomized studies after sufficient adjustment. In other words, they estimated the causal effect of mathematics ($z = 0$) versus vocabulary treatment ($z = 1$) in both the randomized and nonrandomized arms to see whether the estimated effects are the same after adjustment for a set of observed covariates. In their discussion of Shadish et al., Long et al. (2008) reanalyzed these data to estimate a different quantity: the causal effect of mathematics ($z = 0$) versus vocabulary training ($z = 1$) within (a) those who prefer mathematics training and (b) those who prefer vocabulary training. In this current article, we ask another question: What is the causal effect of randomization versus preference for those who prefer mathematics training? Similarly, we estimate the causal effect of randomization versus preference for those who prefer vocabulary training.

Goals of This Article

In summary, there has long been an interest in examining whether effects from rigorous scientific trials generalize to real-world scenarios. One particular dimension of generalizability is whether there is outcome heterogeneity across two settings: that of randomization to treatment versus preference of treatment as it occurs in the real world.

Possibly, psychological characteristics such as motivation and adherence may be enhanced when a preferred treatment is given. These characteristics cannot generally be randomized and may be difficult to measure directly; however, the hybrid randomized and nonrandomized DRPT allow for the unbiased estimation of the causal effect of randomization versus preference. We provide a potential outcomes framework with causal inference for generalizability within the DRPT.

Key Theoretical Background

Potential Outcomes Framework

The potential outcomes framework has been commonly used in experimental design, starting with Splawa-Neyman (1990) in the 1930s, and is often referred to as the Rubin causal model (Rubin, 1974, 1977). Let $w = 1$ for those who are randomized to the randomization arm and $w = 0$ for those who are randomized to the preference arm, $z = 1$ for those who are randomized to or choose Treatment 1 and $z = 0$ for those who are randomized to or choose Treatment 0. In the material that follows, we refer to vocabulary training as Treatment 1 and mathematics training as Treatment 0. Also, let y_1 denote the outcome when given Treatment 1 and y_0 the outcome when given Treatment 0. Finally, we assume there is a vector of covariates $\mathbf{x}$ observed for each individual.

Generally, the goal is to estimate the treatment effect $= E(y_1 - y_0)$ over a given population. The Fundamental Problem of Causal Inference (Holland, 1986) is that it is impossible to observe both y_1 and y_0 for the same person. Instead, we use those who receive the treatment (those with $z = 1$) to estimate $E(y_1 | z = 1)$ and those who receive the control (those with $z = 0$) to estimate $E(y_0 | z = 0)$. A naive estimate of the treatment effect then simply takes the difference between these quantities. However, in general,

$$E(y_1 | z = 1) - E(y_0 | z = 0)$$

is not necessarily equal to $E(y_1 - y_0) =$, since the treatment and control groups may differ from the composition of the given target population. In a randomized experiment where treatment is assigned to every subject by randomization, we at least can obtain an unbiased estimate of the treatment effect, averaged over the population of people in the trial. This is because in a randomized experiment, treatment assignment z is unrelated to all attributes of each person and is consequently independent of the potential responses (y_1, y_0) (Fisher, 1935). In this case, $E(y_1 | z = 1) - E(y_0 | z = 0) = E(y_1 - y_0) =$, where the population is the population represented by the subjects in the trial. We note that the randomized experiment estimates for the people in the randomized trial, but not necessarily for those in the target population of interest, since the sample in the trial is frequently not drawn at random (see, e.g., Marcus, 1997).

In nonrandomized trials, differences between the entire population of interest and the population that receives the treatment can sometimes be attributed to a set of observed covariates $\mathbf{x}$ such that it may be possible to make adjustments to compensate for the imbalances with respect to these covariates. Thus, if selection is on observed covariates $\mathbf{x}$, $E_{\mathbf{x}}\{E(y_1 | z = 1, \mathbf{x}) - E(y_0 | z = 0, \mathbf{x})\}$ is equal to $E(y_1 - y_0) =$ (Rosenbaum & Rubin, 1983). We note that in practice hidden bias due to unobserved covariates is almost always a possibility; that is, we can never know with complete certainty that all bias is due to $\mathbf{x}$. However, even if we cannot be certain about a complete bias removal, it makes sense to apply adjustments to reduce the bias due to observed covariates and to be more cautious about making causal claims.

Using the potential outcomes framework, Shadish et al. (2008) sought to show that $E_{\mathbf{x}}\{E(y_1 | z = 1, \mathbf{x}) - E(y_0 | z = 0, \mathbf{x})\}$ is equal to $E(y_1 | z = 1) - E(y_0 | z = 0) = E(y_1 - y_0) =$, for appropriate methods of adjusting for $\mathbf{x}$; that is, they asked whether in practice nonrandomized studies can actually yield estimates of treatment effects similar to the result obtained from a randomized trial. Long et al. (2008) gave a way to estimate $E(y_1 - y_0 | q = 1)$ and $E(y_1 - y_0 | q = 0)$ where $q = 1$ for those who prefer the treatment and $q = 0$ for those

who prefer the control. In other words, Long et al. estimated the treatment effect within strata defined by preference for the treatment or control.

In the current article, we turn attention to the effect of treatment randomization itself, as compared with treatment preference:

$$\begin{aligned} & \text{(a) } E(y_1|w=1, z=1) - E(y_1|w=0, z=1) \\ & \text{(b) } E(y_0|w=1, z=0) - E(y_0|w=0, z=0), \end{aligned}$$

or (a) the effect of randomization versus preference for those who receive the treatment and (b) the effect of randomization versus preference for those who receive the control. We note that the effect of randomization may be different for the different treatment conditions, and in fact we find this exemplified in the motivating example. The challenge is that we again run into the problem of the Fundamental Problem of Causal Inference: We cannot observe the outcomes under both randomization and preference for the same person: Individuals are assigned to either the randomization group or the preference group, not both. We can observe y_1 for $w=1$ and $z=1$ or for $w=0$ and $z=1$, and we can observe y_0 only for $w=1$ and $z=0$ or for $w=0$ and $z=0$. Thus, we consider four potential outcomes for the DRPT in place of conditioning on w as in a and b above: y_{11} is the potential outcome if a subject is randomized to the treatment, y_{01} is the potential outcome if a subject is randomized to the control condition, y_{10} is the potential outcome if a subject is randomized to the preference arm and chooses the treatment, and y_{00} is the potential outcome if a subject is randomized to the preference arm and chooses the control condition. In terms of the four potential outcomes, the preference effects of interest are given by

$$E(y_{11} - y_{10}) = E(y_{11}) - E(y_{10})$$

for the treatment condition and

$$E(y_{01} - y_{00}) = E(y_{01}) - E(y_{00})$$

for the control condition.

Just as randomization or adjustment for selection bias provides a way around the Fundamental Problem of Causal Inference in the standard setting of treatment–control comparisons, we will use the concept of conditional independence to formalize the notion of how randomization or adjustment works to provide valid causal inference for the effect of randomization versus preference in the DRPT.

Conditional Independence

The concept of conditional independence between random variables derives from probability theory. For a summary on conditional independence and references dating back to the 1950s, see Döhler (1980). We say that A is independent of B given C (in Dawid's, 1979, notation: $A \perp B | C$) if the distribution of A given $(B, C) = (b, c)$ depends only on the value c of C , and not further on the value b of B ; that is, once the value of C is specified, any further information regarding B is irrelevant to uncertainty regarding A .

Dawid (1979) first developed the intuitive concept of conditional independence with its own algebra of formal rules. It is possible to derive many properties of conditional independence by regarding a set of five axioms (e.g., Axiom 1 says $A \perp B | C$ implies $B \perp A | C$; Axiom 2

says $A \perp\!\!\!\perp B \mid A$) as a logical system, rather than using more specific properties of probability distributions (Dawid, 1979).

There has been much interest in applying the notion of conditional independence to causal inference (Dawid, 1984), particularly after the explication of the connections between conditional independence and graphical models (Pearl & Paz, 1987). We say that treatment assignment is strongly ignorable when $(y_1, y_0) \perp\!\!\!\perp z \mid \mathbf{x}$: The potential outcomes y_1 and y_0 are independent of treatment assignment z conditional on a set of covariates. If one can plausibly assume that a set of covariates $\mathbf{x}$ satisfy the statement of strong ignorability in an observational study, then the causal effect of the treatment versus control can be evaluated (Pearl, 2009; Rosenbaum & Rubin, 1983). In this next section, we show how conditional independence can be applied to the DRPT to estimate the causal effect of randomization versus preference.

Causal Inference for the DRPT

In this section, we clarify the causal assumptions and inference for the DRPT using the conditional independence framework. This type of transparency with respect to the form of conditional independence assertions involving potential outcomes and the underlying causal assumptions is essential for valid causal conclusions (Rosenbaum & Rubin 1983). Steyer and colleagues (Steyer, Gabler, von Davier, & Nachtigall, 2000; Steyer, Gabler, von Davier, Nachtigall, & Buh 2000; Steyer, Nachtigall, Wüthrich-Martone, & Krause 2002) pursued a more general approach that also allows for measurement error in potential outcomes, but we follow the simpler Rubin causal model for the purposes of this article.

As already discussed above, depending on the random assignment into the randomized ($w = 1$) or preference arm ($w = 0$) of a DRPT and the subsequent assignment or selection into the treatment ($z = 1$) and control condition ($z = 0$), we get four potential outcomes $y_{zw}(z \in \{0,1\}, w \in \{0,1\})$. However, we only observe y_{11} for subjects in the treatment group of the randomized arm; y_{01} for subjects in the control group of the randomized arm; and the potential treatment and control outcomes, y_{10} and y_{00} , only in the respective group of the preference arm. Thus, only the following four conditional expectations can be inferred from the data: $E(y_{11} \mid w = 1, z = 1)$, $E(y_{01} \mid w = 1, z = 0)$, $E(y_{10} \mid w = 0, z = 1)$, and $E(y_{00} \mid w = 0, z = 0)$. As a consequence, the preference effects we are interested in, that is, $E(y_{11} - y_{10}) = E(y_{11}) - E(y_{10})$ and $E(y_{01} - y_{00}) = E(y_{01}) - E(y_{00})$, cannot directly be estimated (Fundamental Problem of Causal Inference).

An unbiased estimation of the preference effects is only possible if we can reasonably assume that selection into the four groups is ignorable. First, selection into the randomized or preference arm of a DRPT is ignorable, since it is based on random assignment; that is, potential outcomes are independent of assignment w : $(y_{11}, y_{01}, y_{10}, y_{00}) \perp\!\!\!\perp w$ (Ignorability Assumption 1). Second, within the randomized arm selection into the treatment and control groups is once again ignorable due to randomization: $(y_{11}, y_{01}, y_{10}, y_{00}) \perp\!\!\!\perp z \mid w = 1$ (Ignorability Assumption 2). Third, within the preference arm, where subjects select themselves into the treatment or control condition according to their preference, selection is only ignorable if we observe a set of covariates $\mathbf{x}$ such that the potential outcomes $(y_{11}, y_{01}, y_{10}, y_{00})$ are independent of treatment selection z given $\mathbf{x}$: $(y_{11}, y_{01}, y_{10}, y_{00}) \perp\!\!\!\perp z \mid \mathbf{x}, w = 0$, with $0 < P(Z = 1 \mid \mathbf{x}) < 1$ (Ignorability Assumption 3). Note that ignorability with regard to the two potential outcomes (y_{10}, y_{00}) would be sufficient, since the other two potential outcomes are never observed in the preference arm. Assumptions 1–3, together with $\mathbf{x} \perp\!\!\!\perp w$ (which holds due to randomization), imply that $(y_{11}, y_{01}, y_{10}, y_{00}) \perp\!\!\!\perp (z, w) \mid \mathbf{x}$ (Ignorability Assumption 4). This follows directly from the algebra of conditional independence (see Dawid, 1979, Lemma 4.3).

Using these ignorability assumptions, we now show that the preference effects can be obtained from the observed data without any bias. By applying Ignorability Assumption 2 and 1 to the conditional expectations of the randomized arm, we get

$$E(y_{11}|w=1, z=1) = E(y_{11}|w=1) = E(y_{11})$$

and

$$E(y_{01}|w=1, z=0) = E(y_{11}|w=1) = E(y_{01}).$$

That is, the unconditional expectations $E(y_{11})$ and $E(y_{01})$, which are required for estimating the preference effects, can be directly inferred from the randomized arm of a DRPT. Using Ignorability Assumption 3 and 1, we obtain the remaining two expectations $E(y_{10})$ and $E(y_{00})$ from the preference arm by conditioning on $\mathbf{x}$:

$$E_x\{E(y_{10}|w=0, z=1, \mathbf{x})\} = E_x\{E(y_{10}|w=0, \mathbf{x})\} = E(y_{10}|w=0) = E(y_{10})$$

and

$$E_x\{E(y_{00}|w=0, z=0, \mathbf{x})\} = E_x\{E(y_{00}|w=0, \mathbf{x})\} = E(y_{00}|w=0) = E(y_{00}),$$

where E_x denotes the expectation with respect to the covariate distribution of the DRPT population. Under these assumptions, we can estimate unbiased preference effects from observed data, since

$$E(y_{11}|w=1, z=1) - E_x\{E(y_{10}|w=0, z=1, \mathbf{x})\} = E(y_{11}) - E(y_{10}) = E(y_{11} - y_{10})$$

and

$$E(y_{01}|w=1, z=0) - E_x\{E(y_{00}|w=0, z=0, \mathbf{x})\} = E(y_{01}) - E(y_{00}) = E(y_{01} - y_{00}).$$

If the preference effects should be estimated for a different population, that is, different to DRPT population such as the subjects in the treatment or control group (but with the subjects from the randomized and preference arm combined), then the conditional expectations of the randomized arm need also be conditioned on covariates $\mathbf{x}$ and averaged across the corresponding distribution of $\mathbf{x}$. In estimating the preference effects separately for the overall treatment population (i.e., subjects in the treated randomized and preference arm) and the overall control population (i.e., subjects in the control condition of both the randomized and preference arms), as we do in this article, Ignorability Assumption 4 can be relaxed to

$$(y_{11}, y_{10}) \perp w | \mathbf{x}, z=1$$

for estimating the preference effect for treatment subjects and

$$(y_{01}, y_{00}) \perp w | \mathbf{x}, z=0$$

for estimating the preference effect for control subjects.

Thus, we have identified the assumptions necessary for estimating the causal effect of randomization versus preference, namely, the assumption that selection differences between the preference treatment and control groups can be attributed to a set of observed covariates $\mathbf{x}$. And we have provided formal causal inference to show that it is sufficient to adjust the combined randomized and preference treatment arms and, separately, the combined randomized and preference control arms.

We note that our results provide an additional approach to analyzing the DRPT, beyond those approaches given by Shadish et al. (2008) and Long et al. (2008). Shadish et al. adjusted only between the treatment and control arms within the preference arm. Long et al. adjusted between the treatment and control arms within preference strata. In estimating preference effects, we adjusted between randomization and preference arms within the treatment and control groups.

Full Matching on the Propensity Score

The previous section shows that adjustments for $\mathbf{x}$ can be made to estimate the causal effect of randomization versus preference. Generally, those adjustments can be made with covariance adjustment, stratification, or matching (Cochran, 1965). In this article, we use full matching on the propensity score to adjust for characteristics $\mathbf{x}$ that may differ between the randomized and preference populations: The individuals who select a particular treatment are likely different from those randomly assigned to that treatment. The propensity score is defined as the probability of receiving the treatment versus control conditional on a set of observed characteristics $\mathbf{x}$ (Rosenbaum & Rubin, 1983). As discussed further below, in our use of propensity scores, we model the probability of being in the randomized arm (vs. the preference arm) separately for those who receive the treatment and those who receive the control. Matching on the propensity score tends to balance observed covariates between the two groups being compared (Rosenbaum, 2002).

Full matching has been shown to be particularly effective at reducing bias due to the covariates (Ming & Rosenbaum, 2000; Stuart & Green, 2008). Full matching forms a series of small subgroups such that each contains at least one treated and at least one control individual (in our case, at least one randomized individual and at least one preference individual). However, the ratio of treated to control in each subclass can vary (e.g., one subclass may have one treated and five controls, whereas another may have two controls and six treated), and the subclasses are chosen to minimize a global distance measure. Full matching on the propensity score can be operationalized with the *optmatch* package in R (Hansen, 2004; Hansen & Klopfer, 2006; R Development Core Team, 2010).

Propensity scores generally estimate the propensity for treatment ($z = 1$) versus control ($z = 0$). However, in the DRPT, we look at the propensity for randomization ($w = 1$) versus preference ($w = 0$) separately for $z = 1$ and again for $z = 0$. In other words, we match randomized subjects who received the treatment to preference subjects who received the treatment and then match randomized subjects who received the control to preference subjects who received the control.

Illustration

We illustrate the methodology described above using a DRPT of introductory psychology students who are randomized to either a randomization arm of vocabulary versus mathematics training (Treatment 1 vs. Treatment 0) or a preference arm of vocabulary versus mathematics training (Treatment 1 vs. Treatment 0; Shadish et al., 2008). This study

collected both vocabulary and mathematics outcomes as well as the following baseline covariates: vocabulary pretest, mathematics pretest, number of prior mathematics courses, liking for mathematics, math-intensive major, liking for literature, preference for literature over mathematics, math anxiety, extraversion, agreeableness, conscientiousness, emotionality, intellect–imagination, depression, race, age, sex, married, mother’s education, father’s education, college credit hours, American College Test (ACT) comprehensive score, high school grade point average (GPA), and college GPA. The posttest mathematics outcome contained 20 mathematics items (10 presented earlier and 10 new), and the vocabulary outcome contained 30 vocabulary items (15 presented earlier and 15 new).

Because individuals who select a particular training program are likely different from those who are randomized to it, to ensure that the individuals being compared are as similar as possible, we used propensity score full matching to equate the randomized and preference groups. This was done separately for those who received vocabulary training (Treatment 1) and those who received mathematics training (Treatment 0). We describe the process here for those who received vocabulary training; the same process was repeated for those who received mathematics training.

First, a propensity score model was fit among those who received vocabulary training (Treatment 1), predicting being in the randomized arm versus preference arm as a function of the baseline characteristics described above. Each individual’s propensity score was obtained as the predicted probability from this logistic regression: the probability of being in the randomized arm versus the preference arm (among those who received vocabulary training). Missing values in the covariates were handled by doing a simple imputation and including in the propensity score model an indicator of missingness for any variable with more than 5% missing. This will essentially match individuals on the observed values of the covariates as well as on the pattern of missing data. This strategy for dealing with missing values in propensity score analyses was recently used by Haviland, Nagin, and Rosenbaum (2007).

We then used full matching on the propensity score to group randomized and preference individuals into matched sets, such that each matched set contained at least one randomized individual and at least one preference individual. Grouping individuals in this way ensures that we are comparing individuals in the two arms (randomized and preference) with similar propensity scores, and by the properties of the propensity score (Rosenbaum & Rubin, 1983), the individuals will also have similar distributions of the observed covariates.

The matching was generally successful at reducing covariate differences between the randomized and preference arms for both the mathematics training (Treatment 0) and vocabulary training (Treatment 1) groups. For the group that received vocabulary training, the initial difference in propensity scores was 1.0 standard deviation; full matching decreased this to 0.01 standard deviation. Before matching, a number of variables had differences of more than 0.25 standard deviation; after matching, none did. Similar, but less dramatic, reductions in the propensity score differences were seen for the mathematics training group. A few variables, such as the number of prior mathematics courses, whether the student likes mathematics, and whether the student prefers literature over mathematics, had very large differences before matching (over 0.5 standard deviation); the matching was able to reduce these differences somewhat, but some small differences still remained. In particular, although the difference in liking for mathematics was reduced to 0.2 standard deviation, the groups were different on the number of prior mathematics courses by 0.4 standard deviation. Other variables with differences of approximately 0.4 standard deviation after matching were liking for literature, conscientiousness, ACT comprehensive score, high school GPA, and college GPA.

We note that covariates showing imbalances differed according to context. Shadish et al. (2008) looked at the imbalance between the preference vocabulary (Treatment 1) and mathematics training (Treatment 0) groups and found that those in vocabulary training tended to have a higher vocabulary pretest, liked mathematics less, liked literature more, preferred literature more, had fewer mathematics-intensive majors, and had a lower proportion of African American students. Our matching of randomized versus preferred vocabulary training population also showed that those who preferred rather than were randomized to vocabulary training tended to have higher vocabulary pretest scores, preferred literature more, had fewer mathematics-intensive majors, were more agreeable, were more open to experience, were married, and had higher college credit hours and ACT scores. Those who preferred rather than were randomized to mathematics training tended to have lower vocabulary pretest scores, had a higher number of mathematics courses, liked mathematics more, liked literature less, preferred literature less, were more emotional, had a higher proportion of African American students, were more likely to be married, and had higher GPA and ACT scores.

Table 1 gives the observed means of the mathematics outcome for the randomized and preference vocabulary training (Treatment 1) and the randomized and preference mathematics training (Treatment 0). As expected, those who received randomized or preference mathematics training outperformed those who received randomized or preference vocabulary training with respect to the mathematics outcome. However, our primary focus is on the difference between the randomized and preference arms for each of the mathematics training and vocabulary training.

We used the aligned-rank test (Hodges & Lehmann, 1962) to test the null hypothesis of an additive effect θ_0 in a comparison of mathematics outcomes for the matched sets of those who prefer mathematics training (Treatment 0) versus the matched sets of those who were randomized to mathematics training (Treatment 0). The rationale for using the aligned-rank test with full matching on the propensity score is as follows: The aligned-rank statistic can be thought of as a generalization of Wilcoxon's signed-rank statistic for matched sets that are not pairs, but may include variable numbers of controls or treated (Rosenbaum, 2010). For a more extensive discussion of the aligned-rank statistic in propensity score analyses, see Rosenbaum (2002, 2010). The same approach was used to compare mathematics outcomes for matched sets of those who preferred vocabulary training (Treatment 1) versus matched sets of those who were randomized to vocabulary training (Treatment 1). The aligned-rank test is an extension of the signed-rank test to full matching in which the number of controls and treated in each matched set is allowed to vary (Rosenbaum, 2002). This test can be inverted to yield a point estimate, the Hodges–Lehmann estimate (Hodges & Lehmann, 1963), as well as a confidence interval. For a more extensive discussion of this approach, see Rosenbaum (2002).

The Hodges–Lehmann estimate for the randomization versus preference effect for mathematics training ($z = 0$) on the mathematics outcome was -0.71 (effect size = -0.36 , $p = .18$; see Figure 1). A negative effect size means that individuals who preferred rather than were randomized to their treatment performed better. Thus, there is a small to moderate preference effect of mathematics training on the mathematics outcome. The Hodges–Lehmann estimate for the randomization versus preference effect for vocabulary training (Treatment 1) on the mathematics outcome was -0.24 (effect size = -0.12 , $p = .72$). The effect size is very small and most likely reflects random variability rather than an actual difference between randomization and preference.

The small to moderate effect of randomization versus preference for mathematics training (Treatment 0) is consistent with the phenomenon that mathematics phobia is much more

common than fear of vocabulary-related subjects. It seems reasonable that those who prefer mathematics would perform better with respect to a mathematics outcome. However, we must also consider whether the significant preference effect reflects the tendency for mathematics-phobic students to avoid the mathematics training. In this case, the significant preference effect may be due to selection bias, at least in part, rather than a true effect of preference.

Figure 2 gives a plot of the propensity scores for the randomized versus preference populations. We note that there is some nonoverlap of the propensity scores: There are some individuals in the randomized arm with very large propensity scores. Those in the preference arm tend to have lower propensity scores. Thus, the mathematics-phobic students may be poorly represented in the preference arm, contributing to the significant preference effect. Also, those with higher propensity scores tended to have lower mathematics scores, leading to the possibility that the significant preference effect may be due, at least in part, to the nonoverlap. To test the sensitivity of our conclusion to this nonoverlap, we also calculated the preference effect with all individuals with propensity scores of more than 0.90 removed from the data set. The resulting preference effect was almost identical to the estimate given above.

Discussion

Three approaches to analyses of vocabulary (Treatment 1) versus mathematics training (Treatment 0) in a DRPT can provide different but useful information. In this carefully constructed DPRT, Shadish et al. (2008) examined whether the preference arm could yield similar estimates of the efficacy of vocabulary versus mathematics training after appropriate adjustment for selection effects. In their analysis, they assumed SUTVA and, thus, did not consider the existence of preference effects in their analyses. Little, Long, and Lin (2008), in their discussion of Shadish et al., found significant vocabulary versus mathematics training effects within preference strata, that is, within those who preferred vocabulary and within those who preferred mathematics.

The approach in the current article looked at the effect of preference versus randomization within vocabulary training (Treatment 1) and again within mathematics training (Treatment 0) and found one significant small to moderate effect of preference on increasing the mathematics outcome. As we described in the introduction, this question is important for understanding whether causal effects may be extrapolated from RCTs to the more real-world setting in which the preferred treatment is used. In contrast, the parameters estimated by Long et al. (2008)—the effect of vocabulary training versus mathematics training within those who prefer vocabulary training and again within those who prefer mathematics training—has less clinical utility, since people in the real world will not consider both treatments.

The assumptions underlying these three approaches also differ. Long et al. (2008) used principal stratification and instrumental variable-type assumptions to derive estimates. Shadish et al. (2008) used ignorability of the preference vocabulary (Treatment 1) versus mathematics (Treatment 0). The current article derived an ignorability condition such that there is ignorability between the preference and randomized arms for those who got vocabulary training (Treatment 1) and between the preference and randomized arms for those who got mathematics training (Treatment 0). This approach does not necessitate the use of principal stratification due to the timing of the double randomization in the DRPT design.

The three approaches also used different forms of adjustment. Shadish et al. (2008) looked at a variety of adjustment approaches including stratification on propensity scores constructed from all covariates, stratification on propensity scores constructed from covariates of convenience, and linear regression using all covariates. Long et al. (2008) estimated subpopulation means using adjusted and unadjusted maximum likelihood, assuming an additive model for training and preference on outcome. The current article used full propensity score matching for adjustment.

The particular method used for adjustment of course is less important than variations in what is estimated (Cook & Steiner, 2010). However, there are several advantages to using propensity score matching rather than covariance adjustment (Rosenbaum, 2002). Propensity score matching provides balance, on average, for multiple covariates rather than just a few. It permits matching on a unidimensional score that provides a direct assessment of the nonoverlap of covariate distributions, whereas covariance adjustment may give a causal estimate based upon extrapolation that may not hold. For example, we might consider the hypothetical extreme example in which only women receive the treatment and only men are controls. Covariate adjustment would provide a causal estimate that could not be generalized to both men and women. In the current article, we found that there may be nonoverlap with respect to mathematics phobia; that is, we may not be able to generalize the causal estimate of randomization versus preference for those with extreme mathematics phobia, since they tended to have a preference against the mathematics training.

Given the limitations of this approach with respect to the possibility of imperfect implementations of the two randomization involved in an DRPT and the failure to reliably measure all covariates required for establishing strong ignorability, it is natural to ask what can be expected in real-world applications of the proposed approach to examine the causal effect of randomization versus preference. At best, we can calculate an unbiased causal effect that signals psychological differences when receiving treatment assignment by randomization rather than by preference. The causal inference described within this article gives a test of whether causal effects from an RCT can be extrapolated or generalized to the real-world setting in which treatment assignment is not done through randomization. As with most tests of hidden bias, this test cannot tell us whether we have adjusted sufficiently for possible hidden bias. For example, if we find no differences across the randomized and preference arms, it does not guarantee that there is no unobserved covariates for which there is a difference. However, if we find that there is a difference across the randomized and preference arms, this provides valuable information that can help to extrapolate effectiveness from the RCT to a more real-world setting. In this, we are encouraged to use a hybrid randomized and nonrandomized design to further investigate causal mechanisms that explain real-world behaviors.

In the current article, we did not consider nonconsent bias, that is, the bias resulting when not all subjects in a target population consent to enter the DRPT. Marcus (1997) gave methods for adjusting for nonconsent bias in the PRPT. Thus, it would be possible to consider a combined PRPT and DRPT if information is collected from those who do not consent to be in the DRPT. In this case, the ignorability proofs and resulting matching would be more complex, but it would give useful information in the case of substantial nonconsent bias.

The Shadish et al. (2008) vocabulary (Treatment 1) versus mathematics training (Treatment 0) data were carefully constructed to have little noncompliance or missing data. However, noncompliance in DRPTs is an important area for further research. For example, in Long et al.'s (2008) analysis of the Women Take Pride DRPT of group versus self-directed behavioral treatment to enhance women's ability to manage cardiac disease, they found that

those who preferred group treatment were 20.8% more likely to adhere, whereas those who preferred self-directed treatment were 33.6% more likely to adhere. In light of evidence that preference can improve adherence, further work should examine noncompliance within the DRPT. In future work we plan to examine whether preference versus randomization produces less noncompliance and also to consider instrumental variable estimates within propensity-score-matched subsets following the approach given by Marcus and Gibbons (2002).

We also consider that the DRPT may run into problems in view of the fact that randomization works on average, but not always. Thus, the randomized arm as the gold standard for the DRPT may be slightly less than perfect, and the estimate of the effect of randomization versus preference may reflect this random variability. Rubin's (2008) discussion of Shadish et al.'s (2008) DRPT recommends rerandomization of the randomized arm, if possible, in the case where randomization did not balance the Treatment 1 and Treatment 0 arms. When rerandomization cannot take place, he recommends block randomization and adjustment for imbalances in the randomization. This highlights an important limitation of the DRPT: The randomized portion will be balanced across the randomized Treatment 1 and Treatment 0 arms on average. Of course, this limitation applies to interpreting evidence from any RCT; sometimes additional covariate adjustment in addition to propensity score matching may reduce bias.

The significance of the approach in the current article is based upon providing transparency with respect to the causal assumptions and the rigor of conditional independence proofs for formalizing causal inference. The inference for the DRPT is based upon the assumption that treatment selection effects within the preference arm can be attributed to a set of observed covariates $\mathbf{x}$, underlining the need for extensive investigation into plausible covariates and their reliable measurement (Steiner, Cook, & Shadish, 2011; Steiner, Cook, Shadish, & Clark, 2010).

Steiner et al. (2011) concluded that even if all constructs determining the selection process are known, the strong ignorability assumption may still not hold if there is hidden measurement error in the covariates making up the propensity score. Thus, a limitation of the proposed approach in this article is that measurement error in the covariates can result in less bias reduction rather than more.

This limitation does not apply to propensity score matching alone, but would also be a problem for other adjustment methods such as covariance adjustment. However, Steiner et al. (2011) concluded that poorly measured effective covariates still reduce more bias than perfectly measured ineffective covariates. Thus, this limitation may be addressed, in part, by using theory and empirical information to guide which covariates are most crucial to measure accurately. In addition, strong ignorability will be more likely to be satisfied with a large set of covariates covering a range of dimensions as well as different measures within each dimension.

We must also consider the possibility that unobserved bias in the preference arm of the DRPT can always exist. Further research should include sensitivity analyses for assessing the potential impact of bias due to unobserved variables. For example, Gastwirth, Krieger, and Rosenbaum's (2000) sensitivity analysis supposes that hidden bias is due to an unobserved binary covariate and ask what is the largest possible one-sided significance level for the aligned-rank test allowing for the impact of a failure to control for the unobserved covariate. Thus, the significant preference effect that we found could be due to unobserved selection effects. On the other hand, it could be even larger than our estimate. Future

sensitivity analyses might be based upon the assumption that preference cannot lead to worse outcomes.

In conclusion, we might ask how well the approach in this article works compared with reasonable alternative approaches. The DRPT provides a design that allows for the estimation of causal effect of randomization versus preference. The hybrid randomized and nonrandomized nature allows for a rigorous assessment of both internal and external validity. We are unaware of other designs that provide this type of inference, which is partially based upon randomization.

This design also gives a formal approach for studying whether unobserved or poorly measured psychological characteristics such as motivation differ across settings. Another approach for studying characteristics such as motivation that may not be observed explicitly was used by Dynarski (2003). She asked whether financial aid increases college attendance. To answer this question, it would not suffice to compare college attendance across those who did and did not receive financial aid because applying for financial aid reflects motivation for continued education. Dynarski used a change in aid policy to study this question. From 1965 to 1981, the U.S. Social Security Administration provided financial aid for college for the children of Social Security beneficiaries. This approach gives an observational approach for understanding the role of motivation: Students with deceased fathers were much more motivated to attend college during the 1965–1981 period when they received financial aid. However, after the period when the program was eliminated (1982–1983), it can be safely assumed that students with deceased fathers had lower motivation to attend college. Although the change in aid policy was cleverly used by Dynarski to study motivation, the DRPT design gives a more firm causal foundation, since it is partly based upon randomization.

In summary, complex designs such as the DRPT can provide a wealth of information about the effects of treatments, as well as the effect of randomization itself. A better understanding of how effects may vary based on the randomization itself has the potential to impact research across a number of fields, including medicine, education, public policy, and public health. Methods such as that presented here provide a way for researchers to start understanding how participation in a trial may affect the generalizability of those trial results to more general settings, a crucial area for more research.

Acknowledgments

We gratefully acknowledge support from the following sources: Center for Collaborative Inner-City Child Mental Health Services Research (P20 MH085983; principal investigator: M. McKay); Advanced Center on Implementation–Dissemination Science in States for Children and Families (The IDEAS Center; P30 MH090322-01 A1; principal investigators: K. Hoagwood and M. McKay); and Institute for Education Sciences, U.S. Department of Education (R305D100033). Elizabeth A. Stuart received funding from the National Institute of Mental Health (K25MH083846).

References

- Braslow JT, Duan N, Starks SL, Polo A, Bromley E, Wells KB. Generalizability of studies on mental health treatment and outcomes, 1981 to 1996. *Psychiatric Services*. 2005; 56:1261–1268. doi: 10.1176/appi.ps.56.10.1261. [PubMed: 16215192]
- Brewin CR, Bradley C. Patient preferences and randomized clinical-trials. *British Medical Journal*. 1989; 299:313–315. [PubMed: 2504416]
- Cochran WG. The planning of observational studies of human populations. *Journal of the Royal Statistical Society: Series A. General*. 1965; 128:234–266. doi:10.2307/2344179.

- Cook TD, Steiner PM. Case matching and the reduction of selection bias in quasi-experiments: The relative importance of the pretest as a covariate, of unreliable measurement, and of mode of data analysis. *Psychological Methods*. 2010; 15:56–68. doi:10.1037/a0018536. [PubMed: 20230103]
- Coronary Artery Surgery Study (CASS). A randomized trial of coronary artery bypass surgery. Comparability of entry characteristics and survival in randomized patients and nonrandomized patients meeting randomization criteria. *Journal of the American College of Cardiology*. 1984; 3:114–128. doi:10.1016/S0735-1097(84)80437-4. [PubMed: 6361099]
- Dawid P. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B. Methodological*. 1979; 41:1–31. doi:10.2307/2984718.
- Dawid P. Causal inference from messy data. *Journal of the American Statistical Association*. 1984; 79:22–24. doi:10.2307/2288327.
- Döhler R. On the conditional independence of random events. *Theory of Probability and Its Applications*. 1980; 25:628–634. doi:10.1137/1125080.
- Dynarski SM. Does aid matter? Measuring the effect of student aid on college attendance and completion. *American Economic Review*. 2003; 93:279–288. doi:10.1257/000282803321455287.
- Ellenberg SS, Finkelstein DM, Schoenfeld DA. Statistical issues arising in AIDS clinical trials. *Journal of the American Statistical Association*. 1992; 87:562–569. doi:10.2307/2290291.
- Fisher, RA. *The design of experiments*. Hafner; London, England: 1935.
- Flay BR, Biglan A, Boruch RF, Castro FG, Gottfredson D, Kellam S, Ji P. Standards of evidence: Criteria for efficacy, effectiveness, and dissemination. *Prevention Science*. 2005; 6:151–175. [PubMed: 16365954]
- Gastwirth JL, Krieger AM, Rosenbaum PR. Asymptotic separability in sensitivity analysis. *Journal of the Royal Statistical Society: Series B. Statistical Methodology*. 2000; 62:545–555. doi:10.1111/1467-9868.00249.
- Greenhouse JB, Kaizar EE, Kelleher K, Seltman H, Gardner W. Generalizing from clinical trial data: A case study. The risk of suicidality among pediatric antidepressant users. *Statistics in Medicine*. 2008; 27:1801–1813. doi:10.1002/sim.3218.
- Hansen BB. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*. 2004; 99:609–618. doi:10.1198/016214504000000647.
- Hansen BB, Klopfer SO. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*. 2006; 15:609–627. doi:10.1198/106186006X137047.
- Haviland A, Nagin DS, Rosenbaum PR. Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods*. 2007; 12:247–267. doi:10.1037/1082-989X.12.3.247. [PubMed: 17784793]
- Heckman JJ. Micro data, heterogeneity, and the evaluation of public policy: Nobel Lecture. *Journal of Political Economy*. 2001; 109:673–748. doi:10.1086/322086.
- Hodges J, Lehmann E. Rank methods for combination of independent experiments in the analysis of variance. *Annals of Mathematical Statistics*. 1962; 33:482–497.
- Hodges J, Lehmann E. Estimates of location based on rank tests. *Annals of Mathematical Statistics*. 1963; 34:598–611.
- Holland PW. Statistics and causal inference. *Journal of the American Statistical Association*. 1986; 81:945–960. doi:10.2307/2289064.
- Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A. Statistics in Society*. 2008; 171:481–502. doi:10.1111/j.1467-985X.2007.00527.x.
- Janevic MR, Janz NK, Dodge JA, Lin X, Pan W, Sinco BR, Clark NM. The role of choice in health education intervention trials: A review and case study. *Social Science & Medicine*. 2003; 56:1581–1594. doi:10.1016/S0277-9536(02)00158-2. [PubMed: 12614707]
- Levitt SD, List JA. What do laboratory experiments measuring social preferences real about the real world? *Journal of Economic Perspectives*. 2007; 21:153–174. doi:10.1257/jep.21.2.153.
- Little RJ, Long Q, Lin X. Comment [On Shadish, Clark, and Steiner (2008)]. *Journal of the American Statistical Association*. 2008; 103:1344–1346. doi:10.1198/016214508000000995.

- Long Q, Little RJ, Lin X. Causal inference in hybrid intervention trials involving treatment choice. *Journal of the American Statistical Association*. 2008; 103:474–484. doi:10.1198/016214507000000662.
- Macias C, Gold PB, Hargreaves WA, Aronson E, Bickman L, Barreira PJ, Fisher WH. Preference in random assignment: Implications for the interpretation of randomized trials. *Administration and Policy in Mental Health and Mental Health Services Research*. 2009; 36:331–342. doi:10.1007/s10488-009-0224-0. [PubMed: 19434489]
- Marcus SM. Assessing non-consent bias with parallel randomized and nonrandomized clinical trials. *Journal of Clinical Epidemiology*. 1997; 50:823–828. doi:10.1016/S0895-4356(97)00068-1. [PubMed: 9253394]
- Marcus SM, Gibbons RD. Estimating the efficacy receiving treatment in randomized clinical trials with noncompliance. *Health Services and Outcomes Research Methodology*. 2002; 2:247–258. doi:10.1023/A:1020319328212.
- Ming K, Rosenbaum PR. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*. 2000; 56:118–124. doi:10.1111/j.0006-341X.2000.00118.x. [PubMed: 10783785]
- Paradise JL, Bluestone CD, Bachman RZ, Colborn DK, Bernard BS, Taylor FH, Saez CA. Efficacy of tonsillectomy for recurrent throat infection in severely affected children. Results of parallel randomized and nonrandomized clinical trials. *New England Journal of Medicine*. 1984; 310:674–683. doi:10.1056/NEJM198403153101102. [PubMed: 6700642]
- Paradise JL, Bluestone CD, Rogers KD, Taylor FH, Colborn DK, Bachman RZ, Schwarzbach RH. Efficacy of adenoidectomy for recurrent otitis media in children previously treated with tympanostomy-tube placement: Results of parallel randomized and nonrandomized trials. *Journal of the American Medical Association*. 1990; 263:2066–2073. doi:10.1001/jama.1990.03440150074029. [PubMed: 2181158]
- Pearl J. Causal inference in statistics: An overview. *Statistics Surveys*. 2009; 3:96–146. doi:10.1214/09-SS057.
- Pearl, J.; Paz, A. Graphoids: A graph-based logic for reasoning about relevance relations. In: du Boulay, B.; Hogg, D.; Steels, L., editors. *Advances in Artificial Intelligence, II: Seventh European Conference on Artificial Intelligence, ECAI-86; Brighton, U.K.. July 20–25, 1986; Amsterdam, the Netherlands: North-Holland; 1987. p. 307-315.*
- R Development Core Team. *A language and environment for statistical computing*. R Foundation for Statistical Computing; Vienna, Austria: 2010.
- Rosenbaum, PR. *Observational studies*. 2nd ed. Springer-Verlag; New York, NY: 2002.
- Rosenbaum, PR. *Design of observational studies*. Springer; New York, NY: 2010.
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55. doi:10.1093/biomet/70.1.41.
- Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. 1974; 66:688–701. doi:10.1037/h0037350.
- Rubin DB. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*. 1977; 2:1–26. doi:10.3102/10769986002001001.
- Rubin DB. Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*. 2008; 103:1350–1353. doi:10.1198/016214508000001011.
- Rucker G. A two-stage trial design for testing treatment, self-selection and treatment preference effects. *Statistics in Medicine*. 1989; 4:477–485. [PubMed: 2727471]
- Shadish WR, Clark MH, Steiner PM. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*. 2008; 103:1334–1344. doi:10.1198/016214508000000733.
- Shadish, WR.; Cook, TD.; Campbell, DT. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton-Mifflin; Boston, MA: 2002.
- Solomon RL. An extension of control group design. *Psychological Bulletin*. 1949; 46:137–150. doi:10.1037/h0062958. [PubMed: 18116724]
- Splawa-Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*. 1990; 5:465–480. doi:10.2307/2245382.

- Steiner PM, Cook TD, Shadish WR. On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*. 2011; 36:213–236. doi:10.3102/1076998610375835.
- Steiner PM, Cook TD, Shadish WR, Clark MH. The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*. 2010; 15:250–267. doi: 10.1037/a0018719. [PubMed: 20822251]
- Stevens J, Kelleher K, Greenhouse J, Chen G, Xiang H, Kaizar, Arnold LE. Empirical evaluation of the generalizability of the sample from the Multimodal Treatment Study for ADHD. *Administration and Policy in Mental Health and Mental Health Services Research*. 2007; 34:221–232. doi:10.1007/s10488-006-0097-4. [PubMed: 17053977]
- Steyer R, Gabler S, von Davier AA, Nachtigall C. Causal regression models II: Unconfoundedness and causal unbiasedness. *Methods of Psychological Research Online*. 2000; 5(3):55–86. Retrieved from <http://www.dgps.de/fachgruppen/methoden/mpr-online/issue11/art4/steyerCRII.pdf>.
- Steyer R, Gabler S, von Davier AA, Nachtigall C, Buhl T. Causal regression models I: Individual and average causal effects. *Methods of Psychological Research Online*. 2000; 5(2):39–71. Retrieved from <http://www.dgps.de/fachgruppen/methoden/mpr-online/issue10/art3/steyerCRII.pdf>.
- Steyer R, Nachtigall C, Wüthrich-Martone, Krause K. Causal regression models III: Covariates, conditional, and unconditional average causal effects. *Methods of Psychological Research Online*. 2002; 7(1):41–68. Retrieved from <http://www.dgps.de/fachgruppen/methoden/mpr-online/issue16/art3/steyer.pdf>.
- Stuart EA, Green KM. Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*. 2008; 44:395–406. doi:10.1037/0012-1649.44.2.395. [PubMed: 18331131]
- Zelen M. Randomized consent designs for clinical trials: An update. *Statistics in Medicine*. 1990; 9:645–656. doi:10.1002/sim.4780090611. [PubMed: 2218168]

Figure 1.
Full matching for $z = 0$ (mathematics training).

Figure 2. Hodges–Lehmann (H-L) estimates of randomization versus preference of vocabulary training and mathematics training.

Table 1

Observed Means (and Standard Error) for Mathematics Score

<u>$w = 1, z = 1$ ($n = 115$)</u>		<u>$w = 0, z = 1$ ($n = 131$)</u>		<u>$w = 1, z = 0$ ($n = 119$)</u>		<u>$w = 0, z = 0$ ($n = 79$)</u>	
Randomized vocabulary training		Preference vocabulary training		Randomized mathematics training		Preference mathematics training	
<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
7.18	0.3623	7.39	0.3669	11.34	0.2657	12.43	0.3680