Since 1994 Inter-University Consortium
ALMALAUREA
Connecting Universities, the Labour Market and Professionals

AlmaLaurea Working Papers – ISSN 2239-9453

ALMALAUREA WORKING PAPERS no. 9

September 2011

Propensity Score Methods for Causal Inference: On the Relative Importance of Covariate Selection, Reliable Measurement, and Choice of Propensity Score Technique

by

Peter M. Steiner
University of Wisconsin–Madison

This paper can be downloaded at:

AlmaLaurea Working Papers series
http://www.almalaurea.it/universita/pubblicazioni/wp

Also available at:

REsearch Papers in Economics (RePEC)

The **AlmaLaurea working paper series** is designed to make available to a wide readership selected works by AlmaLaurea staff or by outside, generally available in English or Italian. The series focuses on the study of the relationship between educational systems, society and economy, the quality of educational process, the demand and supply of education, the human capital accumulation, the structure and working of the labour markets, the assessment of educational policies.
Comments on this series are welcome and should be sent to pubblicazioni@almalaurea.it.

AlmaLaurea is a public consortium of Italian universities which, with the support of the Ministry of Education, meets the information needs of graduates, universities and the business community. AlmaLaurea has been set up in 1994 following an initiative of the Statistical Observatory of the University of Bologna. It supplies reliable and timely data on the effectiveness and efficiency of the higher education system to member universities' governing bodies, assessment units and committees responsible for teaching activities and career guidance.

AlmaLaurea:

- facilitates and improves the hiring of young graduates in the labour markets both at the national and international level;
- simplifies companies' search for personnel, reducing the gap between the demand for and supply of qualified labour (www.almalaurea.it/en/aziende/);
- makes available online more than 1.5 million curricula (in Italian and English) of graduates, including those with a pluriannual work experience (www.almalaurea.it/en/);
- ensures the optimization of human resources utilization through a steady updating of data on the careers of students holding a degree (www.almalaurea.it/en/lau/).

Each year AlmaLaurea plans two main conferences (www.almalaurea.it/en/informa/news) in which the results of the annual surveys on Graduates' Employment Conditions and Graduates' Profile are presented.

AlmaLaurea Inter-University Consortium | viale Masini 36 | 40126 Bologna (Italy)
Website: www.almalaurea.it | E-mail: pubblicazioni@almalaurea.it

---

International Conference on
"Human Capital and Employment in the European and Mediterranean Area"
Bologna, 10-11 March 2011

Propensity Score Methods for Causal Inference: On the Relative Importance of Covariate Selection, Reliable Measurement, and Choice of Propensity Score Technique

by

Peter M. Steiner*

The author was supported in part by grant R305D100033 from the Institute of Educational Sciences, U.S. Department of Education. I thank Thomas D. Cook, Kelly Hallberg, Steffi Pohl, and William R. Shadish for helpful discussions of studies reviewed in this paper.

Abstract:
The popularity of propensity score (PS) methods for estimating causal treatment effects from observational studies has increased during the past decades. However, the success of these methods in removing selection bias mainly rests on strong assumptions, like the strong ignorability assumption, and the competent implementation of a specific propensity score technique. After giving a brief introduction to the Rubin Causal Model and different types of propensity score techniques, the paper assess the relative importance of three factors in removing selection bias in practice: (i) The availability of covariates that are related to both the selection process and the outcome under investigation; (ii) The reliability of the covariates' measurements; And (iii) the choice of a specific analytic method for estimating the treatment effect—either a specific propensity score technique (PS matching, PS stratification, inverse-propensity weighting, and PS regression adjustment) or standard regression approaches. The importance of these three factors is investigated by reviewing different within-study comparisons and meta-analyses. Within-study comparisons enable an empirical assessment of PS methods' performance in removing selection bias since they contrast the estimated treatment effect from an observational study with an estimate from a corresponding randomized experiment. The empirical evidence indicates that the selection of covariates counts most in reducing selection bias, their reliable measurement next most, and the mode of data analysis—either a specific propensity score technique or standard regression—is of least importance. Additional evidence suggests that the crucial strong ignorability assumption is most likely met if pretest measures of the outcome or constructs that directly determine the selection process are available and reliably measured.

* University of Wisconsin–Madison
e-mail: psteiner@wisc.edu

1. Introduction

The attention paid to causal inference in general and to propensity score methods for estimating causal effects in particular has been considerably increasing over the last three decades. Though randomized experiments are frequently considered as the "gold standard" for causal inference, randomization is frequently not possible due to ethical, administrative, or budgetary reasons. In such cases quasi-experimental methods like regression-discontinuity designs, interrupted time series analysis, instrumental variable approaches, or non-equivalent control group designs might be employed to estimate the treatment effect of an intervention or program (Shadish, Cook & Campbell, 2002). In comparison to a randomized experiment, quasi-experimental methods do not involve randomization but are instead confronted with different types of selection processes. The selection process involved might be known as for regression-discontinuity designs, where subjects get assigned to a treatment and control condition based on a continuous assignment variable and a strict cutoff (Lee & Lemieux, 2009); Or, it might be unknown as it is the case with non-equivalent control group designs where subjects select themselves or are assigned by administrators or third persons into a treatment or control condition (Rosenbaum, 2002, 2009). Studies involving non-equivalent control groups or instrumental variables are also referred to as observational studies.

In observational studies, the problem associated with differential selection into treatment and control conditions is that unadjusted treatment effects are very likely biased. For instance, if unemployed persons most promising for getting a job in the near future are assigned into a labor market program their average employment rate after program participation is very likely higher than the average employment rate of the non-participating unemployed persons—not necessarily because of the treatment effect but because of the better initial position participants were in (this effect is frequently called "creaming"). In such a situation we can estimate unbiased treatment effects only if we are able to adequately model the selection procedure (e.g., Heckman, 1974, 1979), statistically control for observed selection differences (e.g., Cochran & Rubin, 1973; Rubin, 2006), or identify a reliable source of exogenous variation (i.e., an instrument; see Angrist, Imbens & Rubin, 1996)—otherwise bias due to differential selection remains. The assumptions required for an unbiased estimation of causal treatment effects are well known and formulated in statistical theories about causal inference (e.g., Angrist & Pischke, 2009; Rosenbaum, 2002; Rubin, 2006; Steyer, 2005; Steyer et al., 2000a, 2000b) but also in structural causal modeling approaches (e.g., Heckman, 2005; Pearl, 2009).

However, the crucial question in applied research is whether these assumptions required for an unbiased estimation of the treatment effect are actually met for an observational dataset in hand. If the assumptions are not met the estimated treatment effect is very likely biased and the causal claims drawn from the observational study might be invalid. While some of the assumptions involved in a causal inference from observational data are testable like assumptions about the statistical model (e.g., normality or homoscedasticity of the error term) many design assumptions like strong ignorability or exogeneity are not. Unfortunately, effect estimates are frequently much more sensitive to violations of untestable assumptions than testable ones. Thus, violations of untestable assumptions are especially a potential threat to the validity of causal inferences, unless we can convincingly rule them out by carefully chosen design elements like non-equivalent outcome measures or multiple comparison groups (for a discussion of design elements see Shadish, Cook & Campbell, 2002).

In this paper we exclusively focus on propensity score (PS) methods for removing selection bias from observational studies. We investigate empirical evidence about whether PS methods actually work in practice and under which conditions they likely succeed or fail in reducing selection bias. According to theory, PS methods succeed in removing selection bias from observational data if all confounding covariates are reliably measured, the propensity score model is correctly specified, and an appropriated PS technique is chosen. But do these conditions hold in practice? And are some of these conditions more important than others? Using results from within-study comparisons, which

compare effects estimates from PS analyses to comparable benchmark estimates from randomized experiments, the primary objective of this paper is to present empirical evidence with regard to the relative importance of the selection of covariates, their reliable measurement, and the choice of a specific PS method. A secondary objective is to discuss empirical conditions, like the availability of pretest measures of the outcome, for an (almost) complete reduction of selection bias. Using meta-analytical results, we also address the question whether PS methods do better in practice than covariance adjustments via standard regression analysis.

The paper is organized as follows. The next section introduces the potential outcomes notation of the Rubin Causal Model and presents the theoretical assumptions underlying the estimation of causal treatment effects using PS or standard regression techniques. The third section briefly outlines the most frequently used types of PS methods. The fourth section discusses the design and rational of within-study comparisons for evaluating PS methods' success or failure in estimating unbiased causal treatment effects in practice. The review section then present results from different types of within-study comparisons and meta-analyses. Finally, the last section summarizes the findings and concludes with a discussion of the generalizability of the findings.

2. The Rubin Causal Model & Definition of the Treatment Effect

The Rubin Causal Model (Rubin, 1974, 1979; Holland, 1986) with its potential outcomes notation provides a convenient way for formulating the critical assumptions underlying causal inferences from randomized experiments or observational studies (alternative formalizations of causal models can be found in Pearl, 2009, or Steyer et al., 2000a, 2000b). In its simplest formulation with one treatment (Z = 1) and one control condition (Z = 0), the Rubin Causal Model postulates two potential outcomes for each subject $i = 1, ..., N$: a potential control outcome Y_i^0 which is observed if subject i receives the control condition ($Z_i = 0$), and a potential treatment outcome Y_i^1 which is observed if subject i receives the treatment condition ($Z_i = 1$). Y_i^1 and Y_i^0 are called potential outcomes because these outcomes are unknown but fixed prior to treatment selection. Which of the two potential outcomes then actually realizes depends on subject i's selection or assignment to the treatment or control condition. Given the pair of potential outcomes (Y^0, Y^1), different causal estimands of interest can be defined. For illustrative purposes we only define the average treatment effect for the overall target population (ATE) and do not discuss conditional treatment effects like the average treatment effect for the treated (ATT). Using the potential outcomes notation, we can define the average treatment effect (ATE) as the expected difference between potential treatment and control outcomes:

$$\tau = E(Y_i^1 - Y_i^0) = E(Y_i^1) - E(Y_i^0) \qquad (1)$$

However, since we rarely observe both potential outcomes (Y^0, Y^1) simultaneously for each single subject we cannot directly estimate the average treatment effect. This is known a as the "fundamental problem of causal inference" (Holland, 1986). Depending on the treatment status Z, the observed outcome is either the potential control or potential treatment outcome: $Y_i = Y_i^0(1 - Z_i) + Y_i^1 Z_i$ (Rubin, 1974). Since we observe the potential treatment outcomes exclusively for the treatment subjects we can only infer the conditional expectation of treatment outcomes, $E(Y_i \mid Z_i = 1) = E(Y_i^1 \mid Z_i = 1)$, instead of the unconditional expectation $E(Y_i^1)$ which was used in defining the treatment effect in equation (1). Analogously, we only obtain the conditional expectation of control outcomes, $E(Y_i \mid Z_i = 0) = E(Y_i^0 \mid Z_i = 0)$, for the control group. Since these conditional expectations differ in general from the unconditional expectations, $E(Y_i^1)$ and $E(Y_i^0)$,

the prima facie effect, that is the unadjusted difference between the conditional expectations, is in general not equal to the causal estimand defined in (1):

$$E(Y_i^1 \mid Z_i = 1) - E(Y_i^0 \mid Z_i = 0) \neq E(Y_i^1) - E(Y_i^0) = \tau \,. \qquad (2)$$

The prima facie effect is typically biased due to differential selection of subjects into the treatment and control condition. An unbiased treatment effect can be estimated only if we can demonstrate that the selection mechanism (or treatment assignment) that gave rise to the observed data is ignorable (Rosenbaum & Rubin, 1983).

Whether a selection mechanism can be considered as ignorable or not mainly depends on the design of the study under investigation. In a randomized experiment, where subjects are randomly assigned into a treatment and control condition, the assignment mechanism is ignorable. This is so because randomization balances the treatment and control group's characteristics— the two groups are statistically equivalent and, thus, comparable. More specifically, randomization also balances the potential treatment and control outcomes between the treatment and control group. That is, potential outcomes are independent of treatment assignment *Z*, which we can denoted as $(Y^0, Y^1) \perp Z$. Under this independence assumption (i.e., ignorability of treatment assignment) it is easy to show that the conditional expectations of the potential outcomes are identical to the unconditional expectations: $E(Y \mid Z = 1) = E(Y^1 \mid Z = 1) = E(Y^1)$ and $E(Y \mid Z = 0) = E(Y^0 \mid Z = 1) = E(Y^0)$. Consequently, the average treatment effect (ATE) can be obtained from the estimable conditional expectations, $\tau = E(Y \mid Z = 1) - E(Y \mid Z = 0)$, which is identical to ATE in equation (1).

However, since randomization is frequently not possible in practice we need to rely on observational data which typically exhibit selection bias due to differential selection into the treatment and control groups. Consequently, the prima facie effect (the unadjusted mean difference between treatment and control group subjects) is in general biased. Fortunately, we can show that the average treatment effect can be estimated without bias if we are able to adequately measure the selection process such that it is ignorable. This condition is formulated by the crucial strong ignorability assumption which is also called conditional independence or unconfoundedness assumption, or selection on observables (Imbens, 2004; Rosenbaum & Rubin, 1983; see Steyer et al., 2000a and 2000b, for a discussion of different and also weaker assumptions). The strong ignorability assumptions states the following: If we have a set of *p* observed covariates $\mathbf{X} = (X_1, \ldots, X_p)'$ such that potential outcomes (Y^0, Y^1) are independent of treatment selection conditional on **X**, that is,

$$(Y^0, Y^1) \perp Z \mid \mathbf{X}\,, \qquad (3)$$

and if treatment probabilities are strictly between zero and one, $0 < P(Z = 1 \mid \mathbf{X}) < 1$, then the selection mechanism is said to be strongly ignorable (Rosenbaum & Rubin, 1983). If the strong ignorability assumption holds we can show that the average treatment effect (ATE) is the difference in the observable conditional expectations of treatment and control group's outcomes: $\tau = E_{\mathbf{X}}\{E(Y \mid Z = 1, \mathbf{X})\} - E_{\mathbf{X}}\{E(Y \mid Z = 0, \mathbf{X})\}$, which is equivalent to the treatment effect defined in equation (1). The inner expectations refer to the expected potential outcomes for a given group and set of values **X**, the outer expectations average the expected potential outcomes across the distribution of covariates **X**. From a practical point of view, the strong ignorability assumption requires the reliable measurement of all constructs **X** that are simultaneously associated with both treatment status *Z* and potential outcomes (Y^0, Y^1).

Given that the strong ignorability assumption is met, we can estimate unbiased treatment effects using multivariate matching, multivariate stratification, or covariance adjustments via standard regression analysis (Cochran & Rubin, 1973). Note that unbiased effects only result if the method

specific assumptions are also met (e.g., the correct specification of the functional form in a regression analysis). The disadvantage of these methods, particularly of multivariate matching and stratification, is that they may be rather inefficient when the number of covariates is large: The more covariates the less likely it is to find treatment and control cases that are identical (or very similar) on all observed covariates. In order to avoid this "curse of dimensionality" Rosenbaum and Rubin (1983) suggested to use a composite score created from the observed covariates: the propensity score (PS). The propensity score is the conditional probability of belonging to the treatment group given the observed covariates **X**, i.e., $e(\mathbf{X}) = P(Z = 1 \mid \mathbf{X})$. Rosenbaum & Rubin (1983) also proved that if the selection mechanism is ignorable given **X** (see equation 3), then it is also ignorable given the propensity score $e(\mathbf{X})$:

$$(Y^0, Y^1) \perp Z \mid e(\mathbf{X}) . \qquad (4)$$

This means that instead of directly using a set of observed covariates **X** it is sufficient to use the corresponding univariate propensity score for establishing a strongly ignorable selection mechanism.

3. Propensity Score Estimators

If the strong ignorability assumption holds statistical methods that appropriately control for the confounding covariates are at least potentially able to remove all the bias. In the following section we briefly outline four different types of propensity score estimators for removing selection bias. All PS techniques require first the estimation of the unknown propensity score $e(\mathbf{X})$. Given the observed data, we may estimate the PS with binomial regression models, like logit or probit models, but we can also use nonparametric approaches like boosted regression (Berk, 2008; McCaffrey, Ridgeway & Morral, 2004). However, there is not yet enough evidence that nonparametric methods do on average better than parametric binomial regression models (Lee, Lessler, & Stuart, 2009; Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008).

In specifying a PS model, the goal is to balance pretreatment group differences on observed covariates, where balance refers to the equivalence of the joint covariate distributions in the treatment and control group. Achieving balance in observed covariates via PS adjustments basically tries to mimic the balance as established via random assignment—randomization guarantees that the treatment and control groups are statistically equivalent with regard to the observed but also unobserved covariates' distribution. However, in strong contrast to random assignment, a PS can only be modeled with respect to balance in observed covariates—there is no guarantee that a PS also balances unobserved covariates. Given a set of covariates, we can assess the balancing property of the estimated propensity scores with descriptive statistics like the standardized mean difference in covariates (e.g., Rubin, 2001) or significance tests like the *t*-test which tests the mean differences between the treatment and control group (e.g., Ho, Imai, King & Stuart, 2007, with a critical discussion of balancing tests).

Once we estimated a balancing propensity score we may used it in basically four different approaches for removing selection bias from the outcome of interest: (i) PS matching: Individual treatment and control subjects are matched on the basis of the estimated PS, that is, for each treatment subject we try to find a control subject with an identical or at least very similar PS; (ii) PS stratification: Treatment and control subjects are stratified on the estimated PS into a rather small number of homogeneous subclasses (frequently between 5 and 10 strata are used); (iii) Inverse-propensity weighting: In analogy to Horvitz and Thompson's inverse probability weighting (1952), each subject gets a weight derived from the estimated PS which can then be used in a weighted least squares regression, for instance; (iv) Regression estimation using PS related predictors: Dummy variables of PS strata or the cubic polynomial of the PS logit, for instance, might be used as predictors in an outcome regression. Moreover, we may combine all these four basic propensity

score techniques with an additional covariance adjustment within the regression framework. For instance we can combine inverse-propensity weighting and covariance adjustment in a weighted least squares regression; Or, run a regular outcome regression on matched data. The underlying rational for such mixed methods—which are also called "doubly robust" methods—is that the correct specification either of the PS model or the outcome regression model is sufficient for a consistent estimation of the causal treatment effect (Robins & Rotnitzy, 1995; Rubin, 1973; Rubin; 1979). However, it is important to note that if both models are misspecified—which is almost always true in practice—mixed methods might actually perform worse than simple PS analyses without an additional covariance adjustment (Kang & Schafer, 2007).

Though this is only a very brief description of PS techniques, it is sufficient for the purpose of this paper. We refer readers who are interested in a more detailed description and discussion of different propensity score techniques, including mixed methods approaches, to Imbens (2004), Lunceford & Davidian (2004), Morgan & Winship (2007), Rubin (2006), Schafer & Kang (2008), or Steiner & D. Cook (in print). The only additional topic to be briefly addressed here are the differences in finite and large sample properties of different PS methods. For instance, PS stratification removes on average only 90% of the selection bias if five strata based on quintiles are used, that is, 10% of the initial selection bias remains (Cochran, 1968; Rosenbaum & Rubin, 1983); PS matching and inverse-propensity matching suffer from finite sample bias due to the lack of exact matches or the lack of sufficient overlap, respectively (for a discussion see Busso, DiNardo & McCrary, 2009); Or, as a final example, regression estimation using PS-related predictors relies on the correct specification of the functional form between the outcome and the PS (or its logit). In addition to the estimators' (un)biasedness and consistency properties they also differ with regard to their efficiency—which we do not discussing here (e.g., Busso, DiNardo & McCrary, 2009; Imbens, 2004; Schafer & Kang, 2008). Though some estimators have better large sample properties than others there is no unique best estimator for finite samples. Thus, the question of interest addressed in this paper is whether in practice some PS techniques remove regularly more bias than others. An additional question is whether PS methods outperform classical regression analysis. We investigate both questions by reviewing corresponding findings from within-study comparisons and meta-analyses.

4. Strategies for Empirically Evaluating Propensity Score Techniques

As discussed above, the success of propensity score techniques in removing selection bias depends on testable and untestable assumptions. While testable assumptions can be directly probed with the data in hand, untestable assumptions cannot. In an empirical evaluation of PS techniques we are primarily interested in the effect of violating untestable assumptions—because these are the most crucial ones—but we are also interested in the relative importance of biases due to the choice of a specific method (e.g., finite sample bias). In order to investigate the effect of violated assumptions we may run simulation studies where the data generating model—including the true treatment effect—is under the investigator's control. Though such simulation studies are very instructive with respect to a method's sensitivity to deviations from required assumptions, they are not very realistic since the complexity of the simulated data generating mechanism rarely matches up with the complexity of real world processes. Moreover, simulation studies cannot indicate whether assumptions are likely to be met in practice or not.

For this reason, an empirical evaluation of PS techniques or other quasi-experimental methods requires a "real world" setting with real subjects, selection processes, treatments, covariates, and outcomes. However, the problem with such settings is that we rarely know the "true" treatment effect, thus, we lack a benchmark to which a PS-adjusted treatment effect can be compared. So, where can we get a reliable causal benchmark from? From a randomized experiment—which is frequently considered as the best design for a valid causal inference, given its reliable

implementation, of course. In an ideal setting, the randomized experiment and the quasi-experiment are conducted within one single study such that all third-variable confounds like different target populations or differences in treatment implementation and measurement can be ruled out. We call such studies that compare quasi-experimental methods to a randomized experiment "within-study comparisons".

As Cook, Shadish & Wong (2008) discuss, the quality and validity of a within-study comparison depends on several conditions: (i) The randomized experiment which provides the causal benchmark for the comparison need to be of high quality—otherwise it cannot serve as a reliable benchmark for evaluating a quasi-experimental method's performance. (ii) The implementation of the quasi-experiment has to be of high quality, too. There is no point in comparing a poorly implemented quasi-experiment to a perfectly implemented randomized experiment. (iii) The quasi-experiment and the randomized experiment need to be comparable. This means that the two experiments should only differ in their selection process (self- or third-person selection vs. random assignment) but everything else is held constant such that third-variable confounds can be ruled out. (iv) Both studies need to estimate the same causal quantity. If the randomized experiment estimates an intent-to-treat effect (ITT) but the quasi-experiment an average treatment effect (ATE) a comparison of the effect estimates is useless. (v) The analysts of the quasi-experiment should be blinded from the randomized experiment's results. Blinding avoids that the PS analysts try to specify their model with an eye on the randomized benchmark to be replicated. (vi) The criteria for comparing the quasi-experiments' and randomized experiments' result need to be clearly defined. That is, when do we consider the results to be statistically equivalent?

If these six criteria are only poorly met valid conclusions about a quasi-experiments' performance in removing selection bias can hardly be made. The history of within-study comparisons reveals that their quality has considerably increased over the last decades (see Cook, Shadish & Wong, 2008). The first within-study comparisons were conducted in the tradition of LaLonde (1986) who had a randomized treatment and control group and a non-equivalent comparison group. In this design, the outcome mean difference between the randomized treatment group and the randomized control group yields the causal benchmark, whereas the mean difference between the randomized treatment group and the adjusted non-equivalent control group yields the quasi-experimental estimate. We call such studies three-arm within-study comparisons since they employ three groups in their comparison (Figure 1). In contrast, four-arm within-study comparisons are based on four groups: a randomized treatment and control group, and a non-equivalent treatment and comparison group from an observational study (Figure 2). Thus, four-arm within-study comparisons consist of two separate studies with their own treatment group. In order to guarantee comparability of the randomized experiment and the observational study in a four-arm design another randomization step is required: participating subjects need to be randomized into the randomized experiment and the observational study in order to ensure that the underlying target populations are exactly the same. In comparison to three-arm designs, four-arm within-study comparisons have the advantage of holding everything constant except for the assignment mechanism. In particular, third-variable confounds like local differences in target populations or differences in instrumentation (measurement of covariates and outcomes) can be ruled out in a four-arm design.

5. Review of Within-Study Comparisons and Meta-Analyses

5.1. Four-Arm Within-Study Comparisons

Given the higher validity of two recent four-arm within-study comparisons by Shadish, Clark & Steiner (2008) and Pohl et al. (2009) we describe these two studies in more detail and provide additional evidence from two reviews of within-study comparisons (Glazerman, Levy & Myers, 2003, Cook, Shadish & Wong, 2008). The two four-arm within-study comparisons have also the advantage that they where reanalyzed especially with respect to the relative importance of covariate

selection, their reliable measurement and the choice of a specific analytic method (Steiner et al., 2010, 2011; Cook & Steiner, 2010; Cook, Steiner & Pohl, 2009).

Shadish et al. (2008) used a four-arm design in order to investigate whether PS methods can reproduce the results from a randomized experiment. According to Figure 2, they first randomly assigned 445 undergraduate psychology students to a randomized experiment (N = 235) or a quasi-experiment with self-selection (N = 210). In the randomized arm, students were randomly assigned to a mathematics or vocabulary training while students assigned to the quasi-experiment were allowed to choose between one of the two trainings. Not surprisingly, considerably more students choose the vocabulary (N = 131) in order to avoid the mathematics training (N = 79). The short trainings consisted either of learning transforming exponential equations or learning the meaning of special vocables. In estimating the mathematics treatment effect on the mathematics outcome the mathematics group was the treatment group and the vocabulary group served as a comparisons group—and vice versa for the effect of vocabulary training on the vocabulary outcome.

In order to control for the selection bias in the quasi-experiment Shadish et al. carefully measured 23 constructs based on 156 questionnaire items. The 23 constructs belong to five broader construct domains: demographics, proxy-pretests, prior academic achievement, topic preference, and psychological predisposition. The proxy-pretest domain included pretests of the mathematics and vocabulary outcome. These pretests are considered as proxy-pretests because they were measured on a different content and scale. The topic preference domain, which later turned out to contain the most relevant covariates for explaining the selection process and removing selection bias (Steiner et al., 2010), consisted of constructs representing students' liking and preference of mathematics and vocabulary but also students' math anxiety. The prior academic achievement domain covered high school and college scores. The big five personality factors and the Beck depression scale formed the psychological predisposition domain. Given the 23 constructs Shadish et al. measured, all the propensity score methods they applied to the quasi-experimental data yielded a treatment effect comparable to the one obtained from the randomized experiment. They also found that considerable selection bias would have remained if demographical covariates would have been the sole measures available.

In a re-analyses of the Shadish et al. data, Steiner et al. (2010) demonstrated the importance of selecting a set of covariates that establishes strong ignorability. For both trainings and outcomes (mathematics and vocabulary), they showed that proxy-pretests and direct measures of the selection process (i.e., constructs from the topic preference domain) are very effective in reducing selection bias while the other domains were not (demographics, prior academic achievement, psychological predisposition). Indeed, two single constructs (out of 23) would have been sufficient for removing almost all the selection bias. These constructs either directly determined the selection mechanism (liking mathematics or preferring math over literature) or the outcome of interest (proxy-pretests in math and vocabulary). If one would have known these covariates in advance the measurement of two constructs would have been sufficient for removing almost all the selection bias. However, in practice we rarely know for sure which covariates are the crucial ones. Thus, Steiner et al. (2010) also investigated how successful bias reduction would have been if the two most effective constructs indentified for each of the two outcomes would not have been measured. It turned out that the remaining 21 covariates together (without the two most effective ones) would have also removed almost all of the selection bias. This suggests that measuring a broad set of covariates that covers different construct domains with multiple measures is a potential strategy for establishing a strongly ignorable selection mechanism. Such a strategy can be pursued especially when no reliable theories or knowledge about the selection process and the outcome model are available. However, note that even having a multitude of construct domains and multiple measures within domains does not necessarily guarantee that all or most of the selection bias is removed.

The results of Steiner et al. (2010) also indicate that the selection of covariates is much more important than the choice of a specific PS method. The authors analyzed the quasi-experimental

data with three PS techniques (PS stratification, inverse-propensity weighting, PS regression estimation) and a standard regression approach (OLS with all 23 covariates) and found that none of the methods they compared performed uniformly best or worst. Moreover, the variation in bias reduction due to the analytic method chosen was much less than the variation induced by selecting different sets of covariates. Their results clearly show that the selection of covariates is much more important for an unbiased estimation of the treatment effect than the choice of an analytic method (given its competent implementation, of course).

In a second study, Steiner et al. (2011) investigated the effect of unreliable covariate measurement on bias reduction using PS methods. For this purpose, they set up a simulation study based on the Shadish et al. data assuming that all 23 covariates where measured without bias (reliability $\rho = 1$) and that all reliably measured covariates together remove all the selection bias. Then they simulated unreliability in covariate measurements by systematically adding measurement error to each covariate (except for demographical covariates) such the individual covariate reliabilities where stepwise reduced by .1 decrements to $\rho \in \{.9, .8, .7, .6, .5\}$. The results of the simulation study can be summarized as follows: First, if selection and the outcome are determined by latent constructs then measurement error in covariates attenuates a covariates potential to remove bias. Second, the reliable measurement of effective constructs (i.e., constructs that have a high potential to remove selection bias) is more important than the reliable measurement of ineffective constructs (which have no or only a small potential of removing selection bias). There is no need for reliably measuring constructs that do not reduce any bias even if perfectly measured. Third, multiple measures from different construct domains are able to partially compensate for each other's unreliable measurement. Thus, having multiple measures reduces the risk of remaining selection bias not only due to unobserved constructs but also due to imperfect measurements (which can be seen as partially unobserved constructs).

The simulation of Steiner et al. (2011) also allows an assessment of the relative importance of covariate selection, their reliable measurement, and the choice of an analytic method since they run the simulations with three different PS methods (PS stratification, inverse-propensity weighting, PS regression estimation) and a standard regression analysis. Not surprisingly, measuring the constructs that are simultaneously related to both treatment selection and potential outcomes is more important than their reliable measurement. In particular, unreliably measured effective constructs typically remove more bias than perfectly measured ineffective constructs. Then, the reliable measurement of constructs is more important than the choice of a specific analytic method. While measurement error systematically attenuates a covariates potential to remove selection bias, the choice of a specific analytic method did not systematically affect bias reduction. No PS method performed uniformly better or worse than all other methods (PS and standard regression).

Given that all these results rely on a single data set, though with two independent treatments and outcomes, they cannot directly be generalized to other settings. However, Pohl et al. (2009) set out to replicate the Shadish et al. study in Berlin (Germany). They basically used the same design but instead of the vocabulary training they had an English training. In the German study 202 students were randomly assigned into a randomized experiment or a quasi-experiment with self-selection. Those in the randomized experiment then got randomly assigned to the mathematics or English training while students in the quasi-experiment chose their preferred training. It is interesting to note that the selection mechanism in their study differed from the Shadish et al. (2008). Instead of avoiding the mathematics training, students actively chose the training according to their needs. Students tended to chose the topic where they thought they could need more training in order to improve their skills. Moreover, it turned out that the selection process did not result in a selection-biased mathematics outcome—the unadjusted treatment effect of the quasi-experiment was almost the same as the treatment effect from the randomized experiment. Thus, there was basically no bias to be removed. Nonetheless, Pohl et al. (2009) showed that PS and standard regression adjustments did not induce bias in the mathematics treatment effect. For the vocabulary outcome, Pohl et al.

basically replicated the findings by Shadish et al. (2008). Controlling for all covariates, either in a PS or standard regression approach, removed all the bias in the vocabulary outcome. The choice of method did not play an important role but the selection of covariates and their reliable measurement did, as demonstrated in Cook, Steiner & Pohl (2010). Thus, the German within-study comparison replicated the findings of the analyses conducted with the Shadish et al. data: The selection of constructs is the most important factor for a successful removal of selection bias, the reliable measurement of constructs is the next most important factor, and the choice of a specific PS method or regression analysis is of least importance, given its competent implementation. This does, of course, not mean that the choice of a particular method is unimportant. For a given dataset, the choice of an analytic method might actually matter—one method might indicate a significant treatment effect, while another method might suggest an insignificant effect.

5.2. Reviews of Three-Arm Within-Study Comparisons

Though the results derived from the two four-arm within-study comparisons give valuable insights regarding the relative importance of covariate selection, reliable measurement, and choice of method, they are not directly generalizable to more realistic settings like evaluations in the context of labor market programs which typically face more complex selection mechanisms and outcome generating models. However, three-arm within-study comparisons in the tradition of LaLonde (1986) offer some further evidence for the findings derived from the four-arm studies. Glazerman, Levy & Myers (2003) reviewed 12 within-study comparisons on labor market programs and Cook, Shadish & Wong (2008) 12 studies from the other social sciences.

The within-study comparisons that Glazerman et al. (2003) investigated are all on job training programs and their causal effect on subsequent earnings. Not all of these comparisons involved PS techniques—they also used multivariate matching or standard regression approaches. It is interesting that none of the quasi-experimental methods was able to reproduce the treatment effect obtained from the corresponding randomized experiments. However, the averages of the quasi-experimental results and the randomized experiments' results were rather similar. Though the quasi-experimental estimates differed from their experimental benchmark estimates, Glazerman et al. (2003) state that "statistical adjustments, in general, reduced bias, but the bias reduction associated with the most common methods—regression, propensity score matching, or other forms of matching—did not differ substantially". Thus, also in these 12 reviewed three-arm within-study comparisons the choice of a specific analytic method was of least importance. With regard to the importance of covariate selection for removing selection bias, the author concluded that "bias was lower when researchers used measures of preprogram earnings and other detailed background measures to control for individual differences" (p. 86) and that "baseline measures of the outcome are important" for bias reduction (p. 80).

Other than Glazerman et al. (2003), Cook et al. (2008) reviewed a rather heterogeneous set of within-study comparisons coming from different fields of social sciences. Their review of within-study comparisons not only included non-equivalent control group designs but also other quasi-experimental designs like regression discontinuity designs. Cook at al. found that "eight of the comparisons produced observational study results that are reasonably close to those of their yoked experiment, and two obtained a close correspondence in some analyses but not others. Only two studies claimed different findings in the experiment and observational study, each involving a particularly weak observational study. Taken as a whole, then, the strong but still imperfect correspondence in causal findings reported here contradicts the monolithic pessimism emerging from past reviews of the within-study comparison literature" (p. 745). The authors concluded that completely knowing and measuring the selection procedure is one way for establishing an ignorable selection mechanism. It is important to note that this does not only refer to regression discontinuity designs but also to some propensity score studies (e.g., Diaz & Handa, 2006). Cook et al. also

pointed out that having geographically close matches of intact groups (e.g., schools or labor markets), which minimizes the initial bias before individual subjects get matched, helps in reducing selection bias. Further, statistically adjusted treatment effects from observational studies are less successful "(1) when comparison cases are selected from national data sets that differ from the intervention group in population, settings, and measurement; (2) when sample sizes in the control or comparison conditions are modest; and (3) when only demographic variables are available as covariates" (p. 746). Cook et al. also mention the importance of pretest measures of the outcome in combination with locally matched intact groups: "We can also trust estimates from observational studies that match intact treatment and comparison groups on […] pretest measures of outcome" (p. 745). Though, the authors did not directly investigate whether the choice of a specific analytic method matters, there is no indication that PS methods did better or worse than standard regression analyses (Cook & Steiner, 2010).

5.3. Meta-Analysis Comparing PS Methods with Standard Regression Methods

The finding that PS and regression methods do not systematically differ is also supported by two meta-analyses in epidemiology. These meta-analyses investigated studies that estimated the treatment effect of interest using both PS and regression methods. Shah et al. (2005) compared test results of 78 pairs of PS and regression estimates from 43 studies and found that only eight out of 78 (10%) differed significantly. Moreover, they investigated 54 differences in effect sizes (i.e., differences in the natural logarithm of odds or hazard ratios) and found that, "on average, propensity score methods gave an odds or hazard ratio approximately 6.4% closer to unity than traditional regression methods." Shah et al. concluded "that the two methods usually did not differ in the strength or statistical significance of associations between exposures and outcomes" (p. 552). The second study by Stürmer et al. (2006) meta-analyzed 69 studies in epidemiology and found that only 13% of all PS estimates had an effect estimate that differed by more than 20% from conventional regression estimates. Like Shah et al. they "found no empirical evidence […] that PS analyses controlled confounding more effectively than did conventional outcome modeling" (p. 440). Though both meta-analyses do not indicate significant differences between PS methods and regression analyses on average, one has to be cautious in concluding that they do not differ in general. It might be that PS methods do better or worse under certain conditions, e.g., depending on the treatment and control groups' sample size, the initial similarity of the groups, or the presumed complexity of the selection and outcome model. Neither the meta-analyses nor the within-study comparison discussed in this article did systematically investigate conditions under which one method could outperform the other.

6. Discussion

In reviewing two four-arm within-study comparisons, two reviews of 24 three-arm within-study comparisons, and two meta-analyses we demonstrated that, in practice, the selection of covariates is the most important factor for removing selection bias from observational studies. If we fail in measuring constructs that are simultaneously related to treatment selection and potential outcomes, given the other measured covariates, the strong ignorability assumption will not hold and estimates will be biased. The second most important factor for a successful bias reduction is the reliable measurement of constructs. If confounding constructs are unreliably measured bias reduction will be attenuated. It is important to note that reliability only matters if latent constructs instead of observed covariates determine the selection mechanism and the potential outcomes. Finally, the choice of an analytic method turned out to be of least importance in removing selection bias. The within-study comparisons did not indicate any systematic differences between different types of PS

methods nor did they reveal differences between PS and standard regression methods. The latter finding was also supported by two meta-analyses from epidemiology.

Though the empirical evidence gathered so far does not suggest that PS methods are superior to standard regression approaches, they have some design advantages (Rubin, 2007, 2008). First, PS techniques enable the investigator to balance pretreatment group differences without using the outcome. This allows the investigator to be blinded from the outcome measure until he freezes the final PS. The separation of PS estimation and outcome analysis also helps in preserving the type-1 error rate for testing the treatment effect. Second, PS methods offer a natural way for checking the heterogeneity of the treatment and comparison group by comparing the groups' distribution of the PS. When the treatment and comparison group overlap only partially on the propensity score, PS methods typically delete non-overlapping subjects and, thereby, avoid relying on extrapolating treatment effects.

Given the crucial importance of measuring a set of covariates that establishes a strongly ignorable treatment assignment, the question is whether we can identify types of covariates that are in general more successful in reducing selection bias than other types. This review of within-study comparisons indicates that two types of covariates do frequently better than others. First, covariates that directly index the selection mechanism very likely remove a considerable part if not all the selection bias. In the case of administrator assignment, measures on which administrators base their selection of subjects need to be collected (e.g., Diaz & Hand, 2006). If subjects select themselves into the treatment or control condition motivational measures for choosing or avoiding the treatment might successfully remove bias (e.g., Steiner et al., 2010). The second group of covariates consists of pretest measures of the outcome. They frequently reduce a major part of the selection bias because of their presumably high correlation with the outcome of interest. In particular, it is hard to imagine a selection mechanism that strongly affects the potential outcomes but is only weakly related to the corresponding pretest measures. Within the group of pretest measures, we can further distinguish between two types: true pretest measures and proxy-pretest measures. While the former are measured on exactly the same content and scale, the latter typically differ in content but also in their scale of measurement. In general, true pretest measures are preferable to proxy-pretest measures since true pretests are more likely higher correlated with the outcome than proxy-pretests. However, a methodological study by Hallberg, Steiner & Cook (2011) on the effect of retaining instead of promoting poorly performing Kindergarteners demonstrates that proxy-pretest measures can do as well as true pretest measures. In their study, two repeated measures either of the true pretest (achievement score on the same content and scale as the outcome) or the proxy-pretest (teacher assessment) removed almost all selection bias. That is, both the true or proxy-pretests were able to remove nearly as much bias on their own as compared to the case where both types of pretests were combined with the more than 200 available covariates. The study also demonstrates that even if the true and proxy-pretest measures would have not been available the rich set of more than 200 covariates would have removed almost the same amount of selection bias as all covariates and pretest measures together. As already noted in the discussion of the two four-arm within-study comparisons, this finding suggests that a potentially successful strategy for removing selection bias from observational studies is to measure multiple construct domains and multiple constructs within domains. Nonetheless, an informed measurement of selection- and outcome-relevant constructs combined with the additional measurement of a heterogeneous set of constructs is preferable to an uninformed and haphazard collection of data.

Despite the increasing number of within-study comparisons that empirically evaluate the performance of PS methods in removing selection bias one needs to be cautious in generalizing the findings discussed above. The generalizability of the results of the four-arm within-study comparisons is restricted due to their laboratory-like setting (short treatments and not very complex selection processes). The validity of some three-arm within-study comparisons is threatened by third variable confounds like the selection of treatment and comparison groups from different local

and focal populations. Moreover, in many within-study comparisons the treatment effects estimated from observational data did not match the estimates from the corresponding randomized experiments. Despite these shortcomings, the finding that the reliable measurement of confounding constructs matters more than the choice of a specific PS or regression method in removing selection bias was observed across all the studies reviewed. However, for a given data set, the choice of a specific method might actually matter, particularly if one method indicates a significant effect and another one an insignificant treatment effect. In such cases finite sample properties of PS estimators (unbiasdness and efficiency) should not be neglected. Independent of the PS method chosen, the credibility of a causal claim rests mainly on the strong ignorability assumption. Thus, empirical studies using PS or other regression techniques for removing selection bias need to thoroughly justify that the strong ignorability assumption is actually met. Substantive theory, expert knowledge, and pilot studies should be used in warranting an ignorable selection mechanism. If reasonable suspicion about the ignorability assumption being met remains, we better abstain from strong causal claims.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 87*, 328-336.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. An Empiricist's Companion. Princeton: Princeton University Press.

Berk, R. A. (2008). *Statistical Learning from a Regression Perspective.* New York: Springer.

Busso, M., DiNardo, J., & McCrary, J. (2009). Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects. Working Paper. http://www.econ.berkeley.edu/~jmccrary/BDM_JBES.pdf

Cochran, W. G. (1968): The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*, 295-313.

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

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724-750.

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of the pretest as a covariate, unreliable measurement and mode of data analysis. *Psychological Methods, 15*(1), 56–68.

Cook, T. D., Steiner, P. M., & Pohl, S. (2009). Assessing how bias reduction is influenced by covariate choice, unreliability and data analytic mode: An analysis of different kinds of within-study comparisons in different substantive domains. *Multivariate Behavioral Research, 44*, 828-847.

Diaz, J. J., & Handa, S. (2006). An assessment of propensity score matching as a nonexperimental impact estimator. *Journal of Human Resources, 41*, 319–345.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy, 589*, 63-93.

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis. Statistical Methods and Applications*. Thousand Oaks: Sage.

Hallberg, K., Steiner, P. M., & Cook, T. D. (2011). The Role of Pretest and Proxy-Pretest Measures of the Outcome for Removing Selection Bias in Observational Studies. Unpublished Manuscript.

Heckman, J. J. (1974). Shadow prices, market wages, and labor supply. *Econometrica, 42*, 679-694.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometriaca, 47*, 153-161.

Heckman, J. J. (2005). The scientific model of causality. *Sociological Methodology*, 35(1), 1-98.

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

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*, 663–685.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945-970.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*(1), 4-29.

Kang, J., & Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating population means from incomplete data. *Statistical Science, 26*, 523-539.

LaLonde, R. (1986). Evaluating the econometric evaluations of training with experimental data. *American Economic Review, 76*, 604-620.

Lee, D. S., & Lemieux, T. (2009). Regression discontinuity designs in economics. *Journal of Economic Literature, 48*(2), 281-355.

Lee, B., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*, 337-346.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via propensity score in estimation of causal treatment effects: A comparative study. *Statistical Medicine, 23*, 2937-2960.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2009). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9*, 403-425.

Morgan, S. L., & Winship C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge: Cambridge University Press.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge: Cambridge University Press.

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis, 31(4)*, 463-479.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association, 90*, 122-129.

Rosenbaum, P. R. (2002). *Observational Studies* (2nd Ed.). New York: Springer-Verlag.
Rosenbaum, P. R. (2009). *Design Observational Studies*. 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.
Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics, 29*, 185-203.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688-701.
Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74*, 318-328.
Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*, 169-188.
Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge: Cambridge University Press.
Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26*, 20-36.
Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics, 2*, 808-840.
Schafer, J. L., & Kang, J. (2008). Average causal effects from non-randomized studies: A practical guide and simulated example. *Psychological Methods, 13*(4), 279-313.
Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety, 17*, 546-555.
Shah, B. R., Laupacis, A., Hux, J. E., & Austin, P. C. (2005). Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology, 58*, 550-559.
Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association, 103*, 1334-1343.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.
Steiner, Peter M. & David L. Cook (in press). Matching and Propensity Scores. In: Little, T. D. (Ed.), *The Oxford Handbook of Quantitative Methods*.
Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics, 36*(2), 213-236.
Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*(3), 250-267.
Steyer, R. (2005). Analyzing individual and average causal effects via structural equation models. *Methodology, 1*, 39-64.
Steyer, R., Gabler, S., von Davier, A. A., Nachtigall, C., & Buhl, T. (2000a). Causal regression models I: Individual and average causal effects. *Methods of Psychological Research Online, 5*(2), 39-71.
Steyer, R., Gabler, S., von Davier, A. A. & Nachtigall, C. (2000b). Causal regression models II: Unconfoundedness and causal unbiasedness. *Methods of Psychological Research Online, 5*(3), 55-87.
Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., & Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology, 59*, 437-447.

Figure 1. Design of a Three-Arm Within-Study Comparison

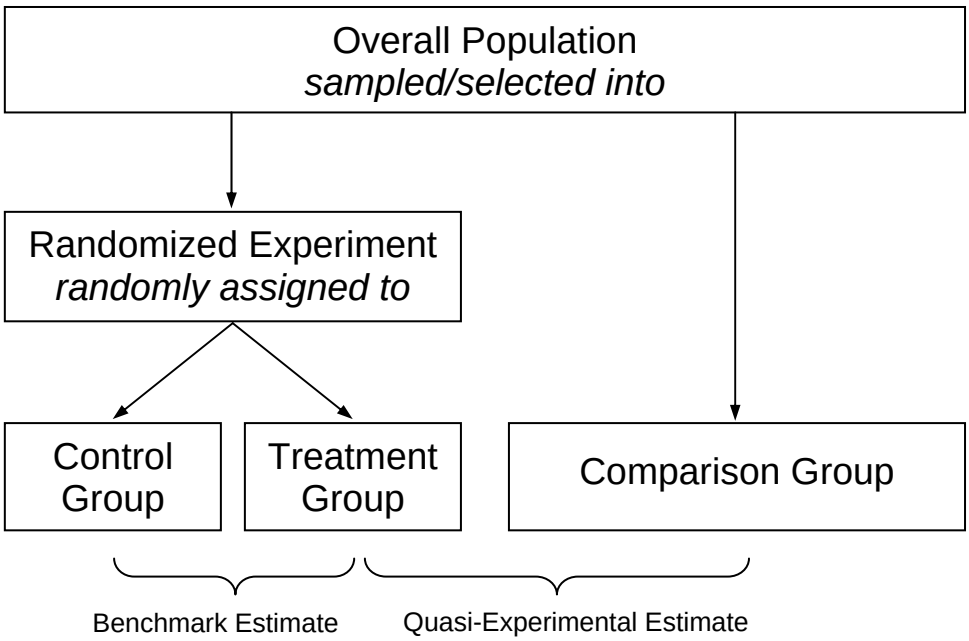

Figure 2. Design of a Four-Arm Within-Study Comparison