Abstract Title Page
Not included in page count.

Title:

On the Importance of Reliable Covariate Measurement in Selection Bias Adjustments Using Propensity Scores

Author(s):

Peter M. Steiner, Northwestern University & Institute for Advanced Studies, Vienna, Austria

Thomas D. Cook, Northwestern University

William R. Shadish, University of California, Merced

Abstract Body

Limit 5 pages single spaced.

Background/context:

Description of prior research and/or its intellectual context and/or its policy context.

In observational studies, causal treatment effects can be estimated without selection bias if the statistical analysis controls for all confounding covariates, i.e., if the strong ignorability assumption is met (Rosenbaum & Rubin, 1983). This entails that (i) all covariates correlated with both treatment selection and the outcome have been measured and that (ii) they are measured without error unless other observed covariates compensate for the unreliable measurement. If there is only a single confounder the effect of its unreliable measurement on the estimated treatment effect is well known (e.g., Cochran, 1968; Campbell and Erlebacher, 1970): An unreliable covariate fails to remove all the selection bias. It is also known that bias is reduced when multiple measures of a crucial variable in the selection model are made, for this reduces unreliability and thus its bias-inducing effects. However, in most propensity score (PS) applications many different measures are used as covariates that were originally collected to represent many different constructs. We typically do not know how well they individually or collectively control for bias and what effects measurement error might have had on the extent of bias reduction achieved, whether for the covariates entering into the final PS selection or for those that we might know to be the most effective in bias reduction.

The best way to learn about effective covariates is to do a within-study comparison in which respondents are first randomly assigned to being in an experiment or observational study and are then randomly or systematically assigned to treatment versus control status. Shadish, Clark & Steiner (2008) did this and showed that all the bias could be reduced by a PS or a simple ANCOVA analysis, and Steiner, Cook, Shadish & Clark (under review) then identified which covariates from among Shadish et al.'s set of 147 items representing 23 constructs from 5 domains were responsible for the bias reduction achieved. So in that one case study we do know which covariates were and were not effective in bias reduction.

This paper investigates how bias reduction was affected when different degrees of measurement error were systematically introduced into the measures constituting the final estimated PS, the PS only for the set of effective covariates and the PS only for the ineffective ones. Since there was already some error in the Shadish et al. covariate measures, a more complex simulation was also done without this source of error. In many ways, this last analysis is the most important.

Purpose/objective/research question/focus of study:

Description of what the research focused on and why.

The study to be reported uses data from a within-study comparison (Shadish et al., 2008, Steiner et al., under review) as the basis for a simulation study. The main purpose of the simulation is to demonstrate how unreliability in covariate measurement affects the degree of bias reduction that would be observed if all the measurement was without error. To this end we systematically reduce the reliability of observed covariates and observe by how much bias reduction is attenuated depending on (i) the degree of measurement error, (ii) the number of

covariates, and (iii) the analytic method used for estimating the treatment effect (PS methods and ANCOVA).

Setting:
Specific description of where the research took place.

The data on which the simulation is built stem from an experiment conducted at a Midsouthern public university.

Population/Participants/Subjects:
Description of participants in the study: who (or what) how many, key features (or characteristics).

445 volunteer undergraduate students from introductory psychology classes.

Intervention/Program/Practice:
Specific description of the intervention, including what it was, how it was administered, and its duration.

445 students were randomly assigned to be in a randomized experiment (N = 235) or a quasi-experiment (N= 210; Figure 1). Those in the randomized experiment were than randomly assigned to participate a vocabulary (N = 116) or mathematics training (N = 119). Those who were assigned to the quasi-experiment chose which training they wanted: 131 chose vocabulary and 79 mathematics. Students in the quasi-experiment attended the same training sessions as those in the randomized experiment. Treatments consisted either of presenting 50 advanced vocabulary terms or five algebraic concepts.

Research Design:
Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

Simulation study based on a within-study comparison of a randomized experiment and quasi-experiment.

Data Collection and Analysis:
Description of plan for collecting and analyzing data, including description of data.

Prior to treatment all students were pretested in vocabulary as well as in mathematics and a rich set of covariates was measured. Overall, 23 covariates based on 147 questionnaire items were used in this study. Covariates belong to 5 domains: Demographics, proxy-pretests, prior-academic achievement, topic preference, and psychological predisposition. After treatment students were assessed in vocabulary and mathematics. In order to estimate causal treatment effects for the vocabulary and mathematics training the mathematics group served as a control

group for the vocabulary group, and the vocabulary group as a control for the mathematics group.

For the simulation study we assumed that all the covariates were measured without error and estimated an ANCOVA model that was then used for creating simulated outcome data. Then we successively reduced the covariates' reliability by adding different amounts of measurement error. We ran a simulation of 2000 replications for three PS methods (stratification, ANCOVA, and weighting) and standard ANCOVA with original covariates. Simulation results were then compared by using linear attenuation rates in treatment effects for different sets of covariates and analytic methods.

Findings/Results:
Description of main findings with specific details.

For a single confounder the attenuation rate in bias reduction is directly related to the reliability of covariate measurement ($0 \leq \rho_X \leq 1$). If the covariate's reliability ρ_X decreases by .1 units 10% less selection bias is removed. Hence, due to the constant attenuation rate measurement error in an effective covariate for bias reduction results in a stronger attenuation of bias reduction (in absolute values) than measurement error in a less effective covariate. Figure 1 represents the results for the vocabulary outcome and Figure 2 for the mathematics one. The first row in each case shows results for the two most effective single covariates for each outcome. There we see the linear attenuation in bias reduction due to variation in the degree of unreliability. The less reliable the measurement of the covariate (ρ_X) the less bias is removed and the more selection bias remains a threat. So the irony is that measurement error has its most negative effect on bias reduction the better the covariate would be if it were measured without error. Differences in analytic methods are quite minor, and mainly due to residual imbalance in specifying the PS models.

The second row shows results for sets of covariates rather than the most single measures most effective in bias reduction. The left-most figures are for all the covariates irrespective of their unique ability to reduce bias. Here we see again that bias is reduced least when measurement error is greatest. The middle plots are for all the covariates, with measurement error introduced only into the most effective single covariates but not the less successful ones. The same result is obtained. As measurement error increases, so does selection bias. However, when the most effective covariates are retained with perfect reliability and only the ineffective covariates are infected with unreliability, then measurement error has no effects on the level of bias reduction achieved. The worse the measure as a construct in a selection model, the less it is affected by how well it is measured. The moral is clear: Measurement error can affect the bias reduction achieved with PS as well as ANCOVA, and the degree of bias reduction suffers when the covariates that would have been most effective when measured without error are in fact measured with error. Measurement error matters in PS work, but only for the better selection constructs.

Conclusions:

Description of conclusions and recommendations of author(s) based on findings and over study. (To support the theme of 2009 conference, authors are asked to describe how their conclusions and recommendations might inform one or more of the above noted decisions—curriculum, teaching and teaching quality, school organization, and education policy.)

The results suggest that the covariates deemed to be most effective in reducing selection bias should be reliably measured. Failure to do so will reduce the bias reduction achieved. In contrast, the reliability of ineffective covariates has minimal effect on bias reduction. Further, the larger the set of interrelated covariates used to control for selection bias the less sensitive is bias reduction to measurement error. But it is crucial to include the singly most effective covariates within this covariate set. We did not find any method specific differences in attenuation rates. When the same covariates are used in all analyses, PS methods show basically the same sensitivity to measurement errors in covariates as ANCOVA,.

Appendixes

Not included in page count.

Appendix A. References

References are to be in APA format. (See APA style examples at the end of the document.)

Campbell, D. T., & Erlebacher, A. E. (1970). How regression artifacts can mistakenly make compensatory education programs look harmful. In J. Hellmuth (Ed.), The Disadvantaged Child: Vol. 3, Compensatory education: A national debate (pp. 185-210). New York: Brunner/Mazel.

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

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

Shadish, W. R., Clark, M. H., and Steiner, P. M. (in press). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association.*

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (under review). The Importance of Covariate Selection in Controlling for Selection Bias in Observational Studies.

Appendix B. Tables and Figures
Not included in page count.

Figure 1. Effect of Measurement Error on Bias Reduction in Vocabulary Treatment Effect by Reliability in Covariates and Analytic Method.

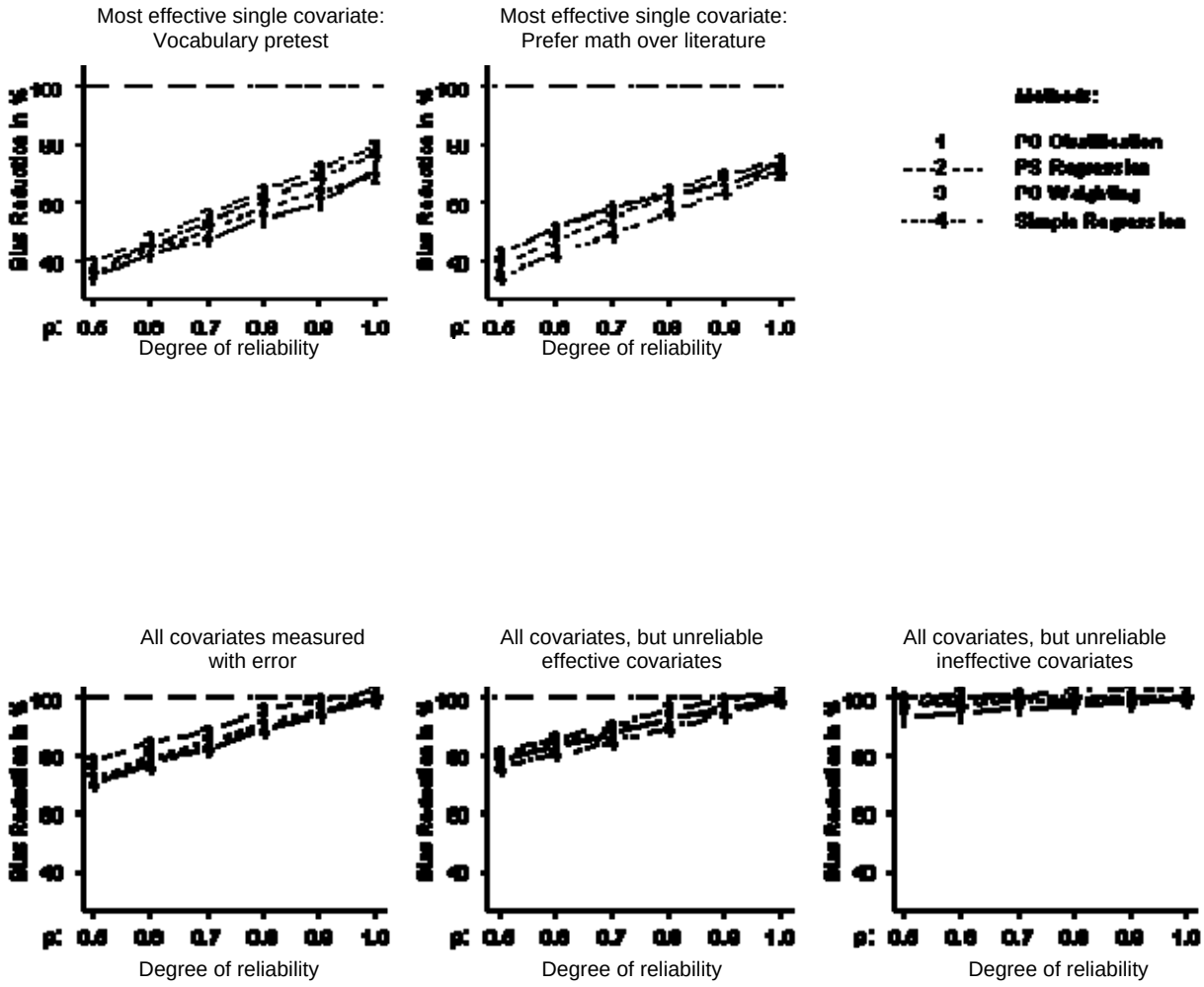

Figure 2. Effect of Measurement Error on Bias Reduction in Mathematics Treatment Effect by Reliability in Covariates and Analytic Method.