Journal of Methods and Measurement in the Social Sciences
Vol.3, No. 2, 1-12, 2012

A Case Study About Why It Can Be Difficult To Test Whether Propensity Score Analysis Works in Field Experiments

William R. Shadish
University of California, Merced

Peter M. Steiner
University of Wisconsin, Madison

Thomas D. Cook
Northwestern University

Peikes, Moreno and Orzol (2008) sensibly caution researchers that propensity score analysis may not lead to valid causal inference in field applications. But at the same time, they made the far stronger claim to have performed an ideal test of whether propensity score matching in quasi-experimental data is capable of approximating the results of a randomized experiment in their dataset, and that this ideal test showed that such matching could not do so. In this article we show that their study does not support that conclusion because it failed to meet a number of basic criteria for an ideal test. By implication, many other purported tests of the effectiveness of propensity score analysis probably also fail to meet these criteria, and are therefore questionable contributions to the literature on the effects of propensity score analysis.

Keywords: Propensity Scores, Strong Ignorability, Quasi-Experiments, Within-Study Comparison

In 2008, Peikes, Moreno and Orzol (henceforth PMO) published a case study in *The American Statistician* that cautioned social program evaluators that propensity score analysis may yield quite different results than those from a randomized experiment. We join them in that caution because we doubt that many applications of propensity scores meet some of the most basic conditions for their valid use (Shadish, 2012; Steiner, Cook & Shadish, 2011; Shadish, Cook, Steiner & Clark, 2010). In that sense, we applaud the PMO article. Field researchers need to appreciate how difficult it can be to use propensity score analysis in a way that yields confidence in the results.

At the same time, however, PMO made a second claim, that they performed an "ideal" (pp. 222, 223, 230) test of whether propensity score matching in quasi-experimental data could approximate the results of a randomized experiment. In fact, a careful analysis of PMO suggests just the opposite conclusion, that they neither implemented an ideal propensity score analysis nor an ideal comparison of results from a propensity score analysis to results from a randomized experiment. In this article, we show why the PMO study was not ideal in both respects. Just as practitioners need to appreciate how difficult it can be to use propensity scores well, methodologists and theorists must appreciate how difficult it is to conduct an ideal test of propensity scores.

In discussing the details of PMO (2008) we rely on their published report plus two detailed earlier reports on the project by Agodini, Thornton, Khan and Peikes (2002) and Peikes, Orzol, Moreno and Paxton (2005). Links to download the latter reports are provided in the reference section. However, to summarize PMO in a nutshell, they first estimated effects in five randomized experiments conducted in three states, New York, New Hampshire, and Oklahoma. Then they replaced the original randomized control group with a nonrandomly formed comparison group constructed through propensity score matching (PSM), and evaluated the comparability of causal estimates between the randomized experiment and the PSM-adjusted nonexperiment. They claimed the estimates were generally different and concluded that the PSM adjusted quasi-experiment failed to identify the correct causal estimate.

Here is the definition of ideal used by PMO: “Our evaluation offered the opportunity to test PSM under seemingly ideal circumstances that included the availability of comprehensive administrative data on a key predictor of both participation and subsequent employment outcomes—employment and earnings for five years before the beginning of the intervention; large pools of potential comparison group members (hereafter candidates); detailed data on program participation; a rigorous protocol for deciding the specification of the propensity score models; and impact estimates derived from experimental methods to validate the performance of PSM” (p. 223). PMO also refer several times to the fact that their database included “hundreds of powerful predictor variables” (p. 229), and that the PSM process passed “multiple statistical tests suggesting that the matching process had worked” (p. 222), where the latter refers to balance tests. In this article, we show that these criteria are neither sufficient for an ideal implementation of propensity score analysis nor for an ideal test of PSM compared to randomized experiments.

Our reasons for rejecting PMO’s (2008) conclusion about how ideal their test was are motivated primarily by six criteria that Cook, Shadish and Wong (2008) have proposed for evaluating comparative studies of the kind that PMO conducted. Hence the bulk of this article analyzes PMO from the perspective of those criteria. At the end of the article, we return to PMO’s definitions of “ideal” and discuss their weaknesses.

Criterion 1.The randomized experiment has to be well-executed if it is to function as a benchmark for validating the results of an adjusted observational study.

High-quality randomized experiments require perfect implementation of randomization, no differential attrition by treatment group, and no treatment noncompliance. Though PMO did not pay explicit attention to the quality of randomization, Peikes et al. (2005) and Agodini et al. (2002) report no pretreatment outcome differences between the treatment and control group, thus suggesting that the random assignment was properly executed with respect to bias in the means. Their use of administrative records suggests that differential attrition due to loss of outcome measurement is unlikely to be a serious problem, though it is clear from Peikes et al. (2005, e.g., Table III.1) that a small amount of overall loss did occur. Although overall loss was only 2%, loss in New York and Oklahoma was 82% and 81%, respectively. In personal correspondence, the authors suggest that these participants or the projects that enrolled them probably submitted incorrect Social Security numbers.

However, treatment noncompliance was considerable and affected standard errors of some of the PMO estimates in a way that may not be properly taken into account in their analyses. Extensive noncompliance occurred in the two New York treatment groups where the notes to PMO's Table 2 indicate that only 29.6% (=277/937) and 32.3% (301/932) of beneficiaries actually complied with their intended treatment. In Oklahoma only 21.8% (314/1,440) did so. Such high noncompliance does not affect the usual intent-to-treat (ITT) estimate. However, the ITT analysis does not estimate the same parameter as the PSM estimate. Consequently, PMO reported a local average treatment effect (LATE) for the randomized experiment computed as the ITT estimate divided by the compliance rate (e.g., Morgan & Winship, 2007). The LATE is typically estimated using an instrumental variable approach with the random assignment indicator as the instrumental variable (IV). In case of one-sided noncompliance, that is, no defiers and no always-takers are present, IV-LATE is equal to the average treatment effect for the treated (TOT; e.g., Angrist, Imbens & Rubin, 1996; Frölich & Melly, 2008). While the absence of defiers and always-takers seems justifiable in PMO's study, it is not clear whether the stable-unit-treatment-value assumption (SUTVA) was actually met. SUTVA would be violated if beneficiaries assigned to the control group in New York tried to compensate for the treatment they knew was withheld from them (Agodini et al., 2002, p. 81). After all, they had been sent a postcard inviting them to hear more about the treatment but then were not given treatment, the kind of circumstance that can lead disappointed control group members to take compensatory actions. But we do know that the low compliance rates function as a weak instrumental variable and generate unreliable estimates and underpowered hypothesis tests. Indeed, all standard errors and *p*-values reported in PMO's Table 2 (except for the small sample New Hampshire site) are underestimated because PMO erroneously used the standard errors of ITT estimates instead of the typically much larger standard errors of IV or Wald estimates. Thus the high noncompliance rates make the PMO randomized experiment a less than ideal benchmark. As a result we are less sure that Table 2 indicates what the population treatment effects really are in the randomized experiment.

Criterion 2. The observational study has to be well done if it is to test the potential of a method as opposed to its robustness under conditions of suboptimal use.

PSM requires two major assumptions: (a) the nonrandomized groups under analysis are balanced and so do not differ by more than chance on the mean and variance of the logit of the estimated propensity score, on all the single covariates composing that score, and on any other pretest variables; and (b) that the strong ignorability assumption holds—i.e., there is no hidden bias from unobservables.

Let us first examine balance. The authors used suboptimal balance tests, relying on the 27 year old Rosenbaum and Rubin (1984) tests rather than more recent ones (e.g., Rubin, 2001) that try to avoid the “balance test fallacy” (Imai, King and Stuart, 2008)—finding non-significant pretest differences that are nonetheless of non-trivial magnitude. The adequacy of balance tests depends on sample size. Sample sizes were adequate in New York and Oklahoma, although we point to different problems with the balance tests for those states shortly. But PMO say explicitly that sample sizes were small in New Hampshire with only 22 treatment and 19 control group participants in the SSI-concurrent experiment and 35 and 34 in the SSDI experiment; however, they still present results for New Hampshire without any further disclaimer that this makes for an extremely poor test of PSM. This small sample size in New Hampshire resulted in very few treated cases in some of the PS-strata used for assessing balance, thus reducing the chance to detect imbalances—some of which were substantial in magnitude despite being nonsignificant. In our view, the results from New Hampshire probably should never have been presented as part of a test of PSM, which is commonly understood to be a large sample method. Such an exclusion would have been consistent with other decisions that PMO made during the design and analysis because of very small sample sizes (Agodini et al., 2002, p. 51; Peikes et al., 2005, p. 52).

The sample size problem also pertains on a reduced scale in New York and Oklahoma, given that samples sizes will inevitably be small in some of the strata used in balance tests. Further, although PMO concluded that their tests indicated substantial overall balance, PMO’s Table 1 shows that this was not the case for a few important covariates. For instance, Table 1 indicates that in the year before enrollment treatment and control employment rates differed by 9.7 (= 41.5 - 31.8) in New York (benefits counseling and waivers) and by -6.1 (= 28.3 - 34.4) percentage points in Oklahoma. This residual imbalance is considerable, given that the differences between PSM and experimental impact estimates for employment rates are of the same magnitude and direction—5.5 (= 14.3 - 8.8) and -6.4 (= 10.6 - 17.0) percentage points, respectively (Table 2 of PMO). These residual imbalances also raise the question whether treatment and comparison groups had enough initial overlap on the propensity score. In case of insufficient overlap bad matches would have resulted because PMO did not delete any treatment participants.

We now turn to the even more crucial strong ignorability assumption. For their covariate choice, PMO relied mostly on administrative records from the Social Security Administration (SSA) but also on administrative records from Census, the Bureau of Labor Statistics, and the Department of Agriculture (Peikes et al., 2005). Such reliance is problematic if the available records do not contain all the covariates necessary for completely modeling selection on variables correlated with outcome. PMO explicitly acknowledge this problem on page 229 or their article, and an earlier report of their study (Agodini et al., 2002, p. 50) states that this condition was not met with the records used: "The sample of potential comparison group members is not limited to those individuals who further met (what we refer to as) the project's secondary criteria, because the SSA data rarely contain information about these criteria. The secondary criteria are more subjective than the primary ones, and include items such as whether it has been determined that a beneficiary needs project services in order to increase earnings substantially" (Agodini et al., 2002, p. 50). Other important variables that the authors say were not in the SSA data were "household composition, occupation, industry and the presence of functional limitations" (p. 55) and "the extent to which individuals are motivated to work and therefore interested in receiving project services" (p. 55). Since variables like need for service, motivation, and functional limitation are plausibly central to the decision to enter the program their omission is nontrivial and calls into question whether strong ignorability was met. Indeed, Agodini et al. (2002, p.16) write: "unmeasured characteristics such as motivation may still bias results". Yet all of these qualifications of the limits of their data base did not stop them from saying they had conducted an ideal test of PSM. PMO also repeatedly stressed that more than 250 covariates were used; but the number of covariates is irrelevant if they fail to tap into some domains non-redundantly responsible for selection.

If PMO had used their analyses to draw conclusions about the inadvisability of propensity scores constructed exclusively from current administrative records, we would perhaps have applauded them. But instead they drew conclusions about the failure of PSM as a general method. Yet neither we nor they know if the method would have failed had if substantive experts and study participants had been polled to suggest what selection processes were likely in the circumstances of the study and if high quality measures of these processes had then been collected (Rubin, 2007; Shadish, 2012; Steiner & Cook, in press). This would probably have pointed to the need to measure a wider range of covariate domains than in PMO, also pushing the authors towards primary rather than secondary data collection in their search for suitable covariates.

Two final issues of technical adequacy also undermine the claim that PMO was an ideal test of PSM. First, Peikes et al. (2005) report missing data in covariates, ranging from 6.8% on type of disability to over 50% for education (Table C.2). To the extent that such data were used to create the propensity scores, results can be quite sensitive to how missing data are handled. PMO apparently used dummy variables to represent missing data, which might be less ideal than modern imputation methods. Using different approaches for handling missing data might help in assessing the results' sensitivity to missing data. Second, PMO used nearest neighbor one-to-one matching, but for many years ideal PSM uses other procedures such as optimal matching (Rosenbaum, 2002). Agodini et al. (2005, p. 66) discuss the merits of the latter, noting it would have increased statistical power and they had developed a computer program to implement it. An ideal test of PSM would have done so.

Criterion 3: There should be no third variables confounding that contrast between the experiment and observational study.

When contrasting experimental and nonexperimental results it is important not to confound the type of causal study with third variables such as location or measurement details. PMO decided to construct a PSM comparison group from a different geographical location than that used in the experiment. This is not ideal. It adds even more uncertainty about potential population differences compared to a comparison group drawn preferably from the same location as the randomized experiment. In a recent review of comparisons of the kind PMO did, Cook et al. (2008) found some evidence that a design using matched local controls may reduce bias better than one using matched non-local controls. This makes sense given that the randomized experiment uses a control group that is by definition from the same location as the treatment group. Ideal PSM would do the same, given that the rationale for PSM is based in trying to reproduce the conditions present in a randomized experiment (Rubin, 2004). It is not sufficient to respond that statistical tests on observed variables showed the nonlocal PSM comparison group to be not different from the randomized group when we already know that the data set may not contain key unobserved variables on which the groups might differ.

Criterion 4. The experiment and nonexperiment should have the same estimator lest differences between estimators masquerade as differences between methods.

In the nonrandomized experiment PMO estimated the average treatment effect for treated (TOT) using a PS model for treatment received. To get a comparable estimate for the randomized experiment they estimated a LATE. Though both TOT and LATE estimate the same causal quantity, LATE relies on some strong additional assumptions (formalized under the IV approached). As argued under criterion one, it is unclear whether these assumptions might have been violated for PMO's randomized experiment, thereby limiting the experiment's role as the criterion for validating PSM. However, PMO could avoid these LATE-related problems by estimating the conventional ITT for the experiment and then by creating an ITT estimate for the nonexperiment by combining the treatment compliers and noncompliers before estimating the propensity score. For PSM, this would require including all beneficiaries assigned to treatment in the randomized experiment and then estimating the propensities for treatment assignment instead of treatment received. Then it is interesting to ask: Would the ITT estimates in the experiment and adjusted nonexperiment be much closer than the corresponding TOT estimates, since the heavy noncompliance in the experiment should radically reduce treatment estimates in the experiment?

Criterion 5. Analysts of the adjusted observational data should be blind to the results of the randomized experiment.

This criterion is designed to protect against analysts inadvertently choosing methods that will bias the nonexperimental design and analysis towards or away from the experimental results. Peikes (personal communication) reported that while the team doing PSM selection was not blinded to the results of the randomized experiment, they were blinded to the post-intervention outcomes of the PSM research sample. This occurred naturally as they needed to complete matching before they obtained follow-up data from the Social Security Administration. Thus PMO probably meets this criterion adequately.

Criterion 6. Defensible standards are used to compare the causal estimates from the experiment and its yoked adjusted nonrandomized experiment.

PMO discuss how much agreement we should expect from comparisons of randomized and adjusted nonrandomized experiments: "If PSM worked, we expected the impact estimates from PSM and the randomized designs to be statistically comparable 95% of the time" (p. 228). Their rationale is that, if the match is good, 5% of any of the estimates should differ from each other by chance, leaving 95% to be comparable. We suspect that this expectation is too high because of implicit assumptions the authors make about studies being implemented identically with perfect power.

There are other ways to set expectations. First, imagine one has two identical replications of the same randomized experiment, each appropriately designed to be powered at .80 to detect a known population effect size. Then the probability that both studies will reject the null is .80 x .80 = .64, and the chances that both will fail to reject the null is .20 x .20 = .04, for a total probability of the same statistical conclusion equal to 68%. The figure drops to 50% if both experiments are underpowered at .50. The figure might fall further if the two randomized experiments are not identical replications in all details; and it might fall even further with every subsequent difference between the randomized and nonrandomized experiments, such as being conducted in different places. So it is not implausible to expect little more than the 27% agreement that PMO actually reported in their Table 2. Of course, if the effect is much bigger than a study was designed to detect, then agreement would occur more often. However, we suspect the latter situation is not characteristic of the comparisons in PMO, given their sample sizes and the large standard errors associated with IV estimates.

Another standard for inferring method differences entails testing whether the estimates from the randomized and the nonrandomized experiments are significantly different from each other. For example, an approximate 95% confidence interval for the mean difference between the randomized mean effect estimate and the PSM mean effect estimate might be given by:

[illegible] ,

where se_p is the pooled standard error of the two estimates. An even more efficient test is obtained when posttest means of the randomized control and PSM comparison group are compared instead of impact estimates (e.g., Bloom et al., 2005). But the more efficient test cannot directly be applied in this case because the randomized control group includes all beneficiaries while the PSM comparison groups were matched to treated participants only (this difference in groups was adjusted by estimating LATE for the randomized experiment). PMO sent us the (probably underestimated) standard errors for the outcome data in their Table 2. This resulted in the observed differences between PSM and the randomized experiment being reliably larger than zero in less than half of the cases (seven of fifteen). In fact, this may underestimate the extent of agreement because the standard errors of their LATE estimates are considerably underestimated (as discussed before). However, this reanalysis is merely an illustration of the more general principle that a proper reanalysis of the raw data could take advantage of better ways to test PSM using correct standard errors for IV estimates.

PMO may correctly respond that policymakers care most about whether the results from the randomized and PSM estimates would lead to the same policy decision, and are relatively little concerned with whether estimates are or are not reliably different from each other. We would agree entirely, but we would also note that the results of studies like PMO are not just of policy interest, and PMO did not just draw policy conclusions. The results are also of scientific interest for statistical theory, and PMO also drew scientific conclusions about the comparability of the estimates. For the latter purpose, the fact that estimates from randomized and PSM adjusted nonrandomized experiments are mostly not significantly different from each other is in accord with our theoretical expectations. For all these reasons, then, the PMO data may actually support the accuracy of the PSM estimates.

Discussion

The six criteria we have invoked above are stringent. Critics may contend that they will lead to comparison studies that can only be done in laboratory settings (e.g., Shadish, Clark & Steiner, 2008). However, we have reviewed recent field-study comparisons of this type (Cook et al., 2008) and found a number that took place in field settings and still were able to remedy many of the problems in PMO (e.g., Aiken, West, Schwalm, Carroll & Hsuing, 1998; Diaz & Handa, 2006). It is challenging, of course, to meet all or most of them, and many practical situations do not allow much more than PMO were able to do. However valorous their study, though, it is not ideal as PMO claimed it to be and so it cannot serve as a strong test of PSM.

This brings us back to the definitions of ideal used by PMO that were cited at the start of this article. The desiderata in those quotes are indeed good things to have when doing PSM. However, they are not enough to result in an ideal test of whether PSM can match the results of a randomized experiment. Examining those desiderata seriatim allows us to quickly summarize many of the problems with the PMO study as a test of PSM (the desiderata from those quotes start each bullet point below):

- “Five years of pretest data on the posttest” and “hundreds of powerful predictor variables” might not be enough for an ideal test when key constructs like need for service, motivation to work, and functional limitation are essential to selection but are missing from the dataset.

- "Large sample sizes from which to construct PS matches" are not enough for an ideal test when the resulting matched comparison group is nonetheless drawn from a different geographical location and so may differ in context specific ways that are not measured and so not used in constructing propensity scores—not to mention the fact that some sites actually had small sample sizes, not large ones.
- "Detailed data on program participation" are not enough for an ideal test when high noncompliance rates make the randomized experiment a less than ideal benchmark in the first place, and when low compliance rates function as a weak instrumental variable that generates unreliable experimental effect estimates.
- "A rigorous protocol for deciding the specification of the propensity score models" is not enough for an ideal test when the protocol itself is suboptimal because the covariates contributing to those models are incomplete, because balance was neither well-tested nor actually achieved on some key pretest covariates, and because better PSM methods could have been used.
- "Impact estimates derived from experimental methods to validate the performance of PSM" are not enough for an ideal test when both attrition and treatment noncompliance in some experimental sites was high, when the assumptions underlying the instrumental variable analysis of the experiment may have been violated, when the experimental standard errors may have been underestimated, and when better estimators in the randomized experiment might well result in effect estimates that are closer to the adjusted quasi-experiment.
- "Multiple statistical tests suggesting that the matching process had worked" are not enough for an ideal test because balance on observed covariates is not sufficient to meet the critical strong ignorability assumption (of course, balance is a precondition for removing selection bias due to observed covariates).

Finally, though not mentioned in their definition of ideal, PMO's criteria for deciding whether the experimental and PSM estimates are similar were also not ideal, or even adequate.

In summary, the PMO study did neither an ideal job of propensity score analysis, nor an ideal job of comparing experimental and PSM estimates. As we said at the start of this article, if the point of the PMO study were simply to caution users of propensity score analysis, we would agree wholeheartedly, though we would be more explicit about the flaws in the PMO propensity score analysis itself. But as a matter of statistical theory as opposed to statistical practice, PMO tells us little or nothing about whether propensity score analysis can work in principle. We can and should do better if we are to construct an empirically based theory of quasi-experimental practice that informs the conditions under which nonrandomized experiments might provide good answers about cause and effect relationships.

Author Notes: William R. Shadish, School of Social Sciences, Humanities and Arts, University of California; Peter M. Steiner, School of Education, University of Wisconsin, Madison; Thomas D. Cook, Institute for Policy Research, Northwestern University. This research was supported in part by grant R305D100033 from the Institute for Educational Sciences, U.S. Department of Education, and by a grant from the University of California Office of the President to the University of California Educational Evaluation Consortium. The opinions expressed are those of the author and do not represent views of the University of California, the Institute for Educational Sciences, or the U.S. Department of Education. Correspondence concerning this article should be addressed to William R. Shadish, School of Social Sciences, Humanities and Arts, University of California, Merced, 5200 North Lake Rd, Merced CA 95343. email: wshadish@ucmerced.edu.

References

Agodini, R., Thornton, C., Khan, N., & Peikes, D. (2002, September). *Design for Estimating the Net Outcomes of the State Partnership Initiative: Final Report* (MPR Reference Number 8663-600). Washington, D.C.: Mathematica Policy Research, Inc. Downloaded from http://mathematica-mpr.com/publications/PDFs/SPIdesign.pdf.

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., & Hsuing, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review, 22*, 207–244.

Angrist, J. D., Imbens, G. W., & Rubin D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*, 444-455.

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*, 724-750.

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

Frölich, M., & Melly, B. (2008). Identification of treatment effects on the treated with one-sided non-compliance. *IZA Discussion Paper No. 3671.*

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

Peikes, D., Moreno, L., & Orzol, S. (2008). Propensity score matching: A note of caution for evaluators of social programs. *The American Statistician, 62*, 222-231.

Peikes, D., Orzol, S., Moreno, L. & Paxton, N. (2005, October). *State Partnership Initiative: Selection of Comparison Groups for the Evaluation and Selected Impact Estimates* (MPR Reference Number 6050-510). Washington, D.C.: Mathematica Policy Research, Inc. Downloaded from http://mathematica-mpr.com/publications/PDFs/SPIselectimpact.pdf.

Rosenbaum, P. R. (2002). *Observational studies* (2nd Ed.). New York: Springer-Verlag.

Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics, 29*, 343-367.

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.

Shadish. W. R. (2012). Propensity score analysis: Promise, reality and irrational exuberance. *Journal of Experimental Criminology, 8,* 1-16

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.

Steiner, P. M., & Cook, D. L. (in press). Matching and Propensity Scores. In T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods*. New York, NY: Oxford University Press.

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*, 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*, 250-267.