A Primer on Propensity Score Analysis☆

William R. Shadish, PhD and Peter M. Steiner, PhD

This article discusses the role that propensity score analysis can play in assessing the effects of interventions. It mostly focuses on identifying the range of solutions to practical problems that occur in propensity score analysis, especially with regard to propensity score construction (logistic regression, classification trees, ensemble methods), balancing (significance tests, other metrics), and analysis (matching, stratifying, weighting, covariance). Throughout, the article will identify particularly important or common pitfalls that need to be avoided in these analyses. The article ends with a discussion of the comparative advantages and disadvantages of propensity scores compared to alternative analytic and design options.

Keywords: Propensity score; Matching; Nonrandomized experiment; Randomized experiment; Strong ignorability

Evidence-based practice is premised on having the capacity to measure accurately the effects of those practices. The strongest case for that capacity is when interventions are assessed with a randomized experiment. When properly implemented and when attrition from conditions is low and not differential, the randomized experiment yields an unbiased and consistent estimate of effects. It is unbiased in that its expectation equals the population parameter, and it is consistent in that any given randomized experiment will converge to its population parameter as sample size increases. The regression discontinuity design, in which participants are assigned to conditions based on whether they fall below or above a cutoff on a measured pretest covariate, has those same characteristics.[1] However, the regression discontinuity design has far less statistical power than the randomized experiment, and so the former is rarely desirable when the latter is feasible.

However, randomized experiments are not always feasible or ethical. In the evaluation of entitlement programs such as Head Start, for example, it was not legal to assign an eligible child to no treatment until the turn of the millennium when Congress specifically mandated such a design. Similarly, researchers cannot randomly assign to conditions when the research is begun after the intervention has already been given to participants, an all too common occurrence. If a regression discontinuity design can be done, it may be the second best choice. Still, the opportunity to assign to conditions based on a cutoff is rare, no doubt partly accounting for the infrequent use of that design until very recently.[2]

When better designs cannot be used, researchers often use an experiment in which participants are not randomly assigned to conditions, sometimes called a quasi-experiment. Here, either participants select their own conditions or a treatment provider or administrator does so. Unfortunately, such methods are presumed to introduce selection bias, raising the threat that the observed effect might be due to differences in participants in the different conditions rather than or in addition to the intervention. Many statistical methods have been proposed to remedy that bias, including selection bias modeling, structural equation modeling, and analysis of covariance. This article discusses one of them, propensity score analysis.[3,4]

For example, consider a nonrandomized experiment with a treatment (coded 1) and a comparison (coded 0) condition. The propensity score is the conditional probability of being in the treatment condition given the set of observed covariates that were measured on participants and/or providers before the start of treatment. Like all probabilities, a propensity score ranges from 0 to 1. The closer it is to 1, the stronger is the prediction that the participant would be in treatment; the closer it is to 0, the stronger the prediction that the participant would be in the comparison condition. In a randomized experiment with equal probability of assigning participants to two conditions, each participant's true propensity score is 0.50. In a nonrandomized experiment, the true probability is unknown for each participant and must be estimated. Participants with similar propensity scores are similar on their overall propensity to be in one condition or the other. Once the propensity scores are created, they can be used in different analyses for equating treatment and comparison conditions on observed pretest variables, with the goal of yielding a better effect estimate.

Following this logic, this article has two major sections. The first section describes how to estimate propensity scores, focusing first on the crucial importance of high-quality measurement of the selection process, then on using those

From the School of Social Science, Humanities and Arts, University of California, Merced, Merced, CA.
From the Institute for Policy Research, Northwestern University, Evanston, IL.
☆ *The authors were supported in part by grant R305U070003 from the Institute for Educational Sciences, U.S. Department of Education. The second author was also supported by grants from the Spencer Foundation and W.T. Grant Foundation.*
Address correspondence to William R. Shadish, PhD, School of Social Science, Humanities and Arts, University of California, Merced, 5200 N. Lake Rd, Merced, CA 95343. E-mails: wshadish@ucmerced.edu, p-steiner@northwestern.edu.

1527-3369/09/1001-0344$36.00/0
doi:10.1053/j.nainr.2009.12.010

measures to create propensity scores with logistic regression or one of several statistical learning methods, and finally on methods for ensuring the propensity scores create a more balanced data set. The second section describes how to use propensity scores to adjust the results of nonrandomized experiments, including matching, stratification, regression, and weighting. The article concludes with a discussion of the strengths and weaknesses of propensity score analysis, including whether it works better than alternatives.

Estimating Propensity Scores

Three issues emerge in estimating propensity scores: the quality of the pretest measures, the statistical method for estimating propensity scores from those measures, and demonstrating balance using the propensity scores.

Quality of Pretest Measures

The most important factor in the successful use of propensity score analysis is the quality of the pretest measures used to create the propensity scores.[5,6] This has a formal statement in propensity score theory: propensity score analysis can produce unbiased estimates of treatment effects if the assumption of strongly ignorable treatment assignment holds.[3] This is the case if treatment assignment (Z_i) and the potential outcomes $Y_i = (Y_{0i},Y_{1i})$ are conditionally independent given the observed covariates $\mathbf{X}_i$ (or, alternatively, the propensity score) that is $\Pr(Z_i|\mathbf{X}_i,Y_i) = \Pr(Z_i|\mathbf{X}_i)$, with $0 < \Pr(Z_i = 1|\mathbf{X}_i) < 1$, where Y_{0i} is the potential comparison outcome (for $Z_i = 0$) and Y_{1i} the potential treatment outcome (for $Z_i = 1$) for all subjects $i = 1,\ldots, N$. These outcomes are called potential outcomes because they refer to the outcome one would observe if subject i gets treated (Y_{1i}) or not treated (Y_{0i}). Note that the potential outcomes are not fully observed because the potential comparison outcomes are only observed for the subjects in the comparison group and the potential treatment outcomes are only observed for those in the treatment group. The ignorability assumption is met if all variables related to both treatment assignment and potential outcomes are included among the covariates (ie, there is no hidden bias) and if there is a nonzero probability of being assigned to the treatment or comparison group for all persons (ie, $0 < \Pr(Z_i = 1|\mathbf{X}_i) < 1$).

The formality of this statistical assumption is sometimes off-putting to many applied researchers. Especially because no empirical test of the assumption exists, the temptation is to simply assume ignorability without further justification for the data being analyzed. The next few paragraphs try to clarify strong ignorability so its practical importance in the conduct of propensity score analysis is better appreciated.

A first crucial matter is ensuring that the available pretest measures assess at least that part of the selection processes that led participants into their respective conditions and is correlated with the potential outcomes. Doing so requires careful and detailed attention to identifying or creating measures of that process. For example, Shadish et al[5] allowed undergraduate students to select themselves into either mathematics or vocabulary training. Before the start of this study, however, they examined the literature, used their own knowledge of undergraduate course preferences, and interviewed student counselors, all to identify or create measures of variables that might influence student choice and correlate with the potential mathematics and vocabulary outcomes—for example, prior mathematics and vocabulary skill levels, whether the student major was mathematics intensive, how many prior mathematics courses they had in high school and algebra, and their expressed liking and preferences for mathematics and literature. For some of these variables, they used existing measures, and for others they created new measures. The need here is for an active investigation into the selection process. Especially to be avoided is the passive acceptance of whatever pretreatment measures may have been gathered or may be available in a data archive. Without specific prior attention to the measurement of selection processes, it is usually very unlikely that commonly available pretreatment measures or data archives include all relevant variables that index selection processes. The one exception may be pretest measurement of the outcome measure, a frequently useful but not always sufficient variable for the construction of good propensity scores. However, without a pretest measure on the outcome it is even harder to justify the strong ignorability assumption.[6,7]

The second crucial aspect of measurement quality is the reliability of the measures which is an issue whenever selection is on latent covariates as is regularly the case with self-selection. Steiner et al[8] showed that the amount of bias reduction obtained from propensity score analysis decreased directly in proportion to the amount of measurement error in the variables used to create the propensity scores—the more measurement error, the less bias reduction. Moreover, this effect was greatest for those variables most directly responsible for bias reduction. If a variable made no contribution to bias reduction, adding measurement error to it had no overall effect on bias reduction.

Given the centrality of quality of measurement for a successful propensity score analysis, a failure to attend to the reliability and validity of measures of the selection process is perhaps the greatest flaw in propensity score analysis today. Common practice seems to be to use whatever pretests measures are available to create propensity scores, with little or no critical analysis of how likely they will reliably assess the selection process. Worse, the standard seems to be to put the most positive light possible on the available covariates, whereas readers deserve to know the shortcomings of a data set for propensity score purposes. If important covariates that are correlated with both treatment and potential outcomes are not measured, the strong ignorability assumption is likely violated and biased effect estimates result.

Statistical Methods for Estimating Propensity Scores

Given the set of measures that the researcher has available, propensity scores are generally estimated using one of two types of methods: binomial regression models or statistical learning

algorithms like classification trees or ensemble methods.[9,10] No matter what method is used, all the available variables that were measured before treatment and are plausibly related to treatment and potential outcomes should be in the initial propensity score model. The most common methods for creating propensity scores are binomial regression models (logistic regression or probit model) in which the propensity to treatment, which is a dichotomous criterion variable, is estimated from the set of pretest measures.[4] Depending on the sample size, the researcher may include all available pretest covariates in the analysis, or start either with a forward or backward stepwise regression when sample size might not support the inclusion of all variables.

The key result from the analysis is a predicted probability of being treated for each person in the sample. That probability is that person's estimated propensity score. However, binomial regression models have a limitation: if the relationship between the pretest measures and treatment assignment is not linear, the researcher must be sure to include appropriate nonlinear terms in the binomial regression model. Otherwise, the resulting propensity scores may not be good estimates of the true propensity scores and fail to achieve good balance on observed covariates. Unfortunately, the functional form of the propensity score might even be highly nonlinear and, hence, hard to address with linear models. Nonlinear regression or generalized additive models might be a more sensitive choice, but they are typically restricted to data with a rather small number of variables.

Statistical learning algorithms including classification tree analysis[11] and ensemble methods[12] like boosting,[13] bagging,[14] or random forests[15] take nonlinearities into account automatically, so that the researcher does not need to guess what the appropriate nonlinear terms might be. Because classification tree methods tend to overfit the data, ensemble methods are usually preferred. The details are not important for present purposes, but to reduce overfitting, each of these methods repeatedly estimates propensity scores on subsets of the data, taking some form of an average or most commonly occurring result as the best propensity score. Given a highly nonlinear selection model, reason exists to think these methods do indeed provide results that more accurately predict treatment condition than binomial regression models.

However, binomial modeling, particularly logistic regression, remains the most commonly used and recommended method, for several reasons. The first reason may not be the best reason—ease of use. Nearly every researcher can easily do a logistic regression to get propensity scores. Learning to do one of the ensemble methods is more challenging, although standard software tools like R strongly facilitate the implementation.[16] But a note of caution is required here because the default settings of corresponding procedures (eg, the depth of the classification tree or the prior distribution of group membership) are frequently not useful with regard to the data on hand. Second, little empirical data exists to suggest that bias reduction is greater using one rather than another of the available methods. Third, if the initial propensity score estimates do not balance pretreatment group differences, it is even less clear than for binomial regression models how to recalibrate the algorithmic procedures to achieve better balance. Fourth, and perhaps most important, the key advantage of the ensemble methods—best prediction of treatment condition membership—may be of less practical interest than might first seem to be the case. To understand why, we turn to the next crucial criterion for propensity score analysis: assessing balance.

Assessing and Obtaining Balance

A mistake made by novice propensity score analysts is to assume that the ultimate goal of the propensity score construction method (eg, logistic regression, bagging) is to optimize an information criterion (eg, Akaike's Information Criterion) or estimate a propensity score that maximizes the correct prediction of the condition each person received—to have the fewest classification errors. Surprisingly, that is not the case, at least not in practice. Indeed, focusing only on the best fit or classification rate may yield less than optimal estimates of the propensity score—that is, a propensity score that does not well balance pretreatment differences in observed covariates between the treatment and comparison conditions.[4,17] Balance is obtained when a propensity score adjustment results in two conditions that have (almost) identical pretest distributions both on the propensity scores themselves and on the available pretest measures. The aim is to mimic the pretest covariate balance obtained in a randomized experiment.

The earliest method for assessing balance advised three steps.[4] First is to divide the participants into five equal strata based on their propensity scores (ie, the lowest 20%, 21%–40%, etc). Second is to cross these five strata with the original treatment-comparison group contrast to yield a 5 × 2 factorial design. Third is to conduct a factorial analysis of variance using this design on each of the pretest covariates. If balance is present, the main effect for treatment and the interaction between treatment and strata should both be nonsignificant across all variables, subject to the usual limitation that 5% of the tests would be significant by chance when tested at a Type I error rate of $\alpha = .05$. The main effect of treatment across all strata is assessed using a weighted average of the stratum-specific differences between the mean of the treatment group and the mean of the comparison group. When all strata have the same number of units, then this is the same as an unweighted average of the five mean differences.

In more recent years, this method has given way to tests of the magnitude of differences between conditions on pretest covariates rather than tests of their statistical significance. Perhaps the most common metric today is a standardized mean bias index. For a continuous variable, for example, one would compute the difference in the treatment and comparison group means, and divide it by an appropriate standard deviation, usually pooled but sometimes just the comparison group standard deviation. This is simply a standardized mean difference statistic (*d*), computed first before propensity score adjustment for assessing the initial imbalance in pretest variables and then afterward for investigating the balance obtained by the propensity score method chosen for estimating

the treatment effect. Several methods can be used, but a common one is to stratify as described in the previous paragraph. Balance is achieved when the index is very close to zero for each of the pretest covariates and also for the propensity score itself. No widely accepted rule exists for how close to zero is needed. Many researchers use $d < 0.20$ or 0.25 but advocate an additional covariance adjustment in the outcome analysis for removing residual bias.[18] Others press to come as close to zero as possible to remove as much bias as possible with the propensity score itself.

Rubin[19] suggests a number of other criteria: (*a*) the standardized difference in the mean propensity score in the two groups should be near 0, (*b*) the ratio of the variance of the propensity score in the two groups should be near 1, and (*c*) the ratio of the variances of the covariates after adjusting for the propensity score must be close to 1, where ratios between 0.80 and 1.25 are desirable, and those smaller than 0.50 or greater than 2.0 are far too extreme. Best practice probably is to use both Rubin's criteria and the *d*-statistic described in the previous paragraph. Rubin[19] also suggests doing all balance and outcome analyses on the logit of the propensity score, which we assume from this point forward.

Commonly, the first set of estimated propensity scores will not achieve the desired level of balance. If so, the researcher should reestimate the propensity scores using a different model and then test for balance again, doing so as many times as it takes either to achieve balance or to conclude that the treatment and comparison group are too different to achieve balance. The propensity score model can be reestimated, for example, by removing nonsignificant covariates, adding previously removed covariates, adding nonlinear functions of predictors, or adding interactions between covariates. The process can be guided by theory about how covariates influence selection, but in the end the process is empirical—the best model is the one that achieves the best balance on all observed variables.

Sometimes the researcher cannot achieve a desirable level of balance on all observed variables. This might be due to a lack of overlap between the treatment and comparison group, indicating that some members of the treatment group show characteristics not found in the comparison group, and vice versa. A clue can sometimes come from a frequency polygon of the propensity score distribution with separate lines for treatment and comparison conditions. For propensity score analysis to be most effective, the two distributions should overlap substantially; the area of overlap is often called the region of common support. When inspection of the distribution shows that the overlap is small, the researcher should not be surprised if balance is difficult or even impossible to achieve. Such cases have an important lesson—some data sets are just not strong enough to support good causal inference on the overall population of interest, and it is a virtue of propensity score analysis that it can suggest such a conclusion before proceeding to the outcome analysis. In such cases, the researcher can delete observations lacking overlap and, if necessary, respecify the propensity score model to achieve good balance with the retained observations. Moreover, in any summary of propensity score analysis results researchers should address the lack of overlap indicating the heterogeneity of groups and the restricted generalizability of estimated treatment effects.

All of the analyses to this point in the article should be done *without looking at the outcome data*.[20] Even better, the researcher creating the propensity scores can be isolated from the outcome data, with the latter analyses done by a second researcher. This helps avoid the danger that the analyst will simply pick the set of propensity scores that most closely indicates the desired conclusion, for example, the set that maximizes the size of the treatment effect.

Using Propensity Scores to Estimate Treatment Effects

Once a best set of propensity scores is identified, those scores are used to estimate adjusted effects from nonrandomized experiments in one of several different ways—matching or stratifying on the propensity score, using the propensity score as a covariate in an ordinary regression, or weighting by some function of the propensity score. All these different propensity score techniques for estimating the treatment effect can be implemented using the regression framework. This has the advantage that the variables already used in creating the propensity score can also be used for an additional covariance adjustment. Indeed, doing so typically reduces residual bias due to imperfect balance or a misspecified propensity score model and improves power by reducing standard errors.[5,21,22] Here we describe how to do such analyses, and the advantages and disadvantages of each, although little evidence exists that one analytic method is routinely superior to another. But first, consider the two different kinds of effects that can be estimated using propensity scores.

What Quantities Are Being Estimated: Average Treatment Effect and Treatment Effect on the Treated

In an ordinary randomized experiment, the outcome analysis compares the treatment group mean to the comparison group mean, perhaps adjusted for some measured covariates. This estimates the average treatment effect (ATE), the effect of treatment across the entire population of treated and untreated units. Contrast this with a different estimator, the average treatment effect on the treated (TOT), estimated only for those who actually received the treatment. In a randomized experiment, those exposed to treatment are probabilistically equivalent to those not exposed to treatment, so ATE and TOT are identical. In nonrandomized experiments, however, those taking treatment may be different in a variety of ways from those not exposed, and so ATE and TOT differ in general. All the analytic methods that follow can be used for estimating both ATE and TOT, although matching methods typically focus on TOT. It is crucial that the analyst knows which quantity is being estimated because the actual estimation procedure depends on the quantity of interest.

Matching

Matching aims to create similar groups of treatment and comparison units by matching together units that have (almost) identical propensity scores.[23,24] The simplest form of matching, one-to-one matching, is best known but perhaps not the best choice. In one-to-one matching without replacement, a treatment participant is paired with a comparison group participant who has the most similar propensity score. Similarity is typically defined as falling within a certain range (caliper), for example, within ±.05 propensity score points. Those two participants are removed from the pool, and then the next treatment participant is matched to the next most similar comparison group participant. This continues until all treatment group participants have a match or until no further matches within the caliper are possible. At that point a comparison of the two group means (eg, using a matched-pairs *t* test) yields the propensity score–adjusted TOT effect estimate. Alternatively, the treatment effect for the matched data may be estimated using a standard regression analysis that also enables the inclusion of additional covariates.

Although simple to do, one-to-one matching typically omits a large number of comparison participants that would have been the second- or third-best matches, for instance. An alternative is to allow each treatment participant to have more than one comparison group match. When the pool of comparison group participants is relatively large compared to the number of treatment group participants, this method can greatly increase statistical power compared to one-to-one matching. Better still is full matching.[23,25] Matched sets are created that may contain multiple treatment or multiple comparison participants, or both. Those sets are chosen with an iterative procedure that aims to minimize the overall propensity score difference between treatment and comparison groups. Once the best matched sets are identified, ATE or TOT can be estimated. Which causal quantity is estimated depends on the choice of weights derived from the matching structure.[25,26] Typically, all forms of matching discard cases that do not belong to the common support region (as defined in the section on balance). This may both reduce power and limit generalization

Stratification

We described how to stratify on propensity scores at the start of the section on balancing tests. The same procedure is followed during outcome analysis except that this time it is done on the outcome variable rather than the pretest covariates. Depending on the stratum weights, both ATE and TOT can be estimated. If ATE is the estimand of interest, stratum weights are determined by the number of treated and untreated in each stratum. For TOT, weights are derived from the distribution of the treated across strata. Following Cochran,[27] using five strata is common because they capture approximately 90% of the variability in propensity scores that would have been captured by an exact matching of the same individuals. Especially when both sample sizes and the region of common support are large, however, more strata can be supported. Hong and Raudenbush,[28] for example, used 7 to 10 propensity score strata. In studies with partially nonoverlapping treatment and comparison groups, strata with either no treatment or no comparison cases may result. Solutions to the latter include dropping the stratum with the empty cell, or changing the boundaries between strata so that no stratum has empty cells. The former reduces power and generalizability of the treatment effect, just as in the case of matching. The stratification approach also allows for an additional covariance adjustment. This can be done by running a regression analysis separately within each stratum and then pooling the regression-based treatment effects as before.

Regression Estimation: Propensity Scores as a Covariate

A third analytic option is to enter the propensity score as a covariate in a regression equation predicting outcome from treatment.[5] Doing so is probably the simplest option, and no data need be eliminated because of poor overlap, thereby retaining the generalizability of results to the target population of interest. However, one obtains unbiased estimates only if the functional form is correctly specified for both the treatment and comparison group. This is particularly difficult if the logit of the propensity score is nonlinearly related to the outcome. And there is no guarantee that including appropriate nonlinear transformations of the propensity scores and interactions with the treatment is sufficient for getting the functional form right. Corresponding results are always subject to doubt about whether this has been done correctly. Hence, regression approaches that are less sensitive to functional form assumptions were suggested.[22,29] These approaches rely on cubic spline regression or dummies based on a categorized propensity score. In addition, they estimate a regression model for each group separately, then predict for each subject both potential outcomes from the treatment and comparison regressions, respectively, and finally compute the treatment effect as the average difference between the predicted potential treatment and comparison outcomes. This procedure also has the advantage that ATE and TOT are precisely defined and estimable. Moreover, all regression methods also allow for an additional covariance adjustment by including covariates already used for estimating the propensity score. The hope with such "doubly robust" procedures is that residual bias due to a misspecified propensity score model is removed via covariance adjustment.[21]

Weighting

Finally, propensity scores can be used to create weights.[30,31] Depending on the causal quantity of interest, different weighting schemes are used. For estimating ATE, the weights are the inverse of the propensity scores for the treatment group, and the inverse of one minus the propensity score for the comparison group. The procedure is to multiply each participant's weight and outcome score, sum those over all participants, and then divide that by the sum of all the weights.

Alternatively, weights can be used in a weighted least squares regression together with an additional covariance adjustment. The procedure is basically the same as described for the regression estimates above, except that weighted least squares is used instead ordinary least squares.[22] A problem with this method is that extreme propensity scores (either very close to zero or one) can result in very large weights that dominate the analysis disproportionately. Robins et al[31] suggest an adjusted weighting method for reducing this influence, or extreme propensity scores can be Winsorized to a less extreme value. If TOT is the estimand of interest, the corresponding estimate is obtained by weighting treatment participants with one and comparison participants by the propensity score divided by one minus the propensity score.

Discussion

Propensity score analysis is increasingly popular for adjusting nonrandomized experiments. But does it work? The answer may be a qualified yes.[5] The crucial qualification was mentioned previously—high-quality measurement of the selection process is crucial. Propensity score analysis is likely to work best when careful measurement of the selection process is planned before the study begins. This should include interviewing potential participants to find out what factors they consider in deciding whether to seek the intervention, and doing similar interviews with providers and managers who might influence who receives the intervention, as well as studying any pertinent research literature. Then, the researcher either can identify existing measures of those selection processes or can create new ones. The process is likely to work less well, or possibly not at all, when the researcher relies on an existing data set that was gathered with no particular concern for good measurement of the selection process. The process is likely to fail when all that is available is routine demographic measures such as ethnicity, sex, or marital status, for it is rare that those measures account for all of the selection process.

Does propensity score analysis work better than alternatives like standard regression analysis with nonrandomized data? The jury is still out on this question, but some evidence suggests that the answer is a qualified no. For example, Shadish et al[5] found that entering all the pretest covariates into a regression equation without any use of propensity scores worked at least as well as propensity score analysis. Indeed, within-study comparisons that compare the results of a randomized experiment to corresponding results of a nonrandomized experiment[32,33] and meta-analyses[34,35] regularly show that PS estimates do not significantly differ from regression estimates. However, the qualifications to this point are not trivial. The first qualification is that propensity score analysis may be preferable when the researcher wishes to match or stratify on a large number of pretest covariates. Matching participants on more than just a few covariates is logistically hard. Propensity scores reduce all the covariates to a single number, which makes matching simple. A second qualification is that the use of the balance tests described earlier in this article can be helpful in diagnosing whether a given data set can support causal inference. Although balance on observed covariates is not sufficient for bias reduction (because all covariates needed for establishing strong ignorability have to be measured, too), a failure to achieve balance or a considerable lack of overlap is a strong signal that the researcher should be cautious about pursuing causal statements.

Is propensity score analysis sufficiently well developed that researchers can confidently use it as an alternative to better designs such as the randomized experiment or the regression discontinuity design? Here the answer is a definite no, for several reasons. First, the number of assumptions that must be made during a propensity score analysis far exceeds those needed for better quality experimental designs like the randomized experiment or the regression discontinuity design. For instance, neither of the latter methods requires as many assumptions about knowledge of the selection process as does propensity score analysis, and this is crucial for obtaining unbiased effect estimates. Second, there is little evidence about how well propensity/PASW score methods perform when the treatment and comparison groups only overlap on the tails of their propensity score or covariate distribution. Third, no generally accepted criteria exist about how much balance is sufficient. For instance, a residual imbalance of 0.2 standard deviations in a covariate that is highly correlated with potential outcomes might cause considerable bias in the treatment effect. Fourth, we still know little about the empirical conditions under which propensity score analysis yields reliable effect size estimates. For example, little work has been done comparing the relative effectiveness of matching, stratifying, covariance, and weighting as propensity score analysis techniques to allow us to say with confidence which of them works best under which conditions, for example, when sample sizes are small or the pool of comparison cases is not considerably larger than the number of treatment cases. Fifth, even under the best conditions, the appropriate techniques for estimating accurate standard errors (eg, bootstrapping) are still under discussion.[18]

For those who wish to use propensity score analysis, all the procedures described in this article can be done using commonly available statistical packages like SAS or SPSS/PASW. Both packages have logistic regression procedures that can produce propensity scores (usually called predicted probabilities in those procedures). All the balancing tests are simple to program with a few lines of syntax. For both packages, different macros for propensity score matching are available (eg, greedy matching[36] or optimal matching[37] in SAS and greedy matching[38] in SPSS) In addition, the statistical software tools Stata and R offer more specialized packages for propensity score analyses (eg, PSMATCH2,[39] MATCH,[40] or PSCORE[41] in Stata, and MatchIt,[42] Matching,[43] or optmatch[44] in R).

Propensity score analysis is a major contribution both to the theory and the practice of effect size estimation in nonrandomized experiments. Yet like any tool, its existence tempts researchers to use it in place of more careful attention to the design of high quality randomized and nonrandomized experiments.[32] Propensity score analysis is no panacea. As we

and others have written elsewhere, in the contest between design and analysis, design rules.[20,45]

References

1. Shadish WR, Cook TD, Campbell DT. Experimental and ouasi-experimental designs for generalized causal inference. Boston: Houghton-Mifflin; 2002.
2. Cook TD. "Waiting for life to Arrive": a history of the regression-discontinuity design in psychology, statistics and economics. *J Econometrics*. 2008;142:636-654.
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.
4. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity scores. *J Am Stat Assoc*. 1984;79:516-524.
5. Shadish WR, Clark MH, Steiner PM. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *J Am Stat Assoc*. 2008;103:1334-1343.
6. Steiner PM, Cook TD, Shadish WR, et al. The importance of covariate selection in controlling for selection bias in observational studies. Psychol Methods. In press.
7. Cook TD, Steiner PM. 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. In press.
8. Steiner PM, Cook TD, Shadish WR. On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. J Educ Behav Stat. In press.
9. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer; 2001.
10. Berk RA. Statistical learning from a regression perspective. New York: Springer; 2008.
11. Stone RA, Obrosky DS, Singer DE, et al. Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia. *Medical Care*. 1995;33:AS56-AS66.
12. Berk RA. An introduction to ensemble methods for data analysis. *Sociol Methods Res*. 2006;34:263-295.
13. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods*. 2004;9:403-425.
14. Luellen JK, Shadish WR, Clark MH. Propensity scores: an introduction and experimental test. *Evaluation Review*. 2005;29:530-558.
15. Berk R, Li A, Hickman LJ. Statistical difficulties in determining the role of race in capital cases: a re-analysis of data from the state of Maryland. *J Quant Criminol*. 2005;21:365-390.
16. R Development Core Team. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing 3-900051-07-0; 2009. http://www.R-project.org.
17. Shadish WR, Luellen JK, Clark MH. Propensity scores and quasi-experimentation. In: Bootzin RR, McKnight P, editors. Strengthening research methodology: psychological measurement and evaluation. Washington (DC): American Psychological Association; 2006. p. 143-157.
18. Ho D, Imai K, King G, et al. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Anal*. 2007;15:199-236.
19. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv Outcome Res Methodol*. 2001;2:169-188.
20. Rubin DB. For objective causal inference, design trumps analysis. *Ann Appl Stat*. 2008;2:808-840.
21. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *J Am Stat Assoc*. 1995;90:122-129.
22. Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychol Methods*. 2008;13:279-313.
23. Rosenbaum PR. Observational studies. 2nd ed. New York: Springer-Verlag; 2002.
24. Rubin DB. Matched sampling for causal effects. New York: Cambridge University Press; 2006.
25. Hansen BB. Full matching in an observational study of coaching for the SAT. *J Am Stat Assoc*. 2004;99:609-619.
26. Stuart EA, Green KM. Using full matching to estimate causal effects in non-experimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Dev Psychol*. 2008;44:395-406.
27. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 1968;24:295-313.
28. Hong G, Raudenbush SW. Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educ Eval Policy Anal*. 2005;27:205-224.
29. Little RJA, An H. Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*. 2004;14:949-968.
30. Hirano K, Imbens G, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*. 2003;71:1161-1189.
31. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11:550-560.
32. Cook TD, Shadish WR, Wong VC. Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. *J Policy Anal Manag*. 2008;27:724-750.
33. Glazerman S, Levy DM, Myers D. Nonexperimental versus experimental estimates of earnings impacts. *Ann Am Acad*. 2003;589:63-93.
34. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol*. 2005;58:550-559.
35. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. 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. *J Clin Epidemiol*. 2006;59:437-447.
36. Parsons LS. Reducing bias in a propensity score matched-pair sample using greedy matching techniques. SAS Institute Inc., Proceedings of the Twenty-Sixth Annual SAS Ò Users Group International Conference, Paper 214-26. Cary, NC: SAS Institute Inc.; 2001. http://www2.sas.com/proceedings/sugi26/p214-26.pdf.
37. Kosanke J, Bergstralh E. Match cases to controls using variable optimal matching. http://mayoresearch.mayo.edu/mayo/research/biostat/upload/vmatch.sas. 2004.
38. Levesque R. Raynald's SPSS Tools. http://www.spsstools.net/Syntax/RandomSampling/MatchCasesOnBasisOfPropensityScores.txt.
39. Leuven E, Sianesi B. PSMATCH2. Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical Software Components S432001. Chestnut Hill, Massachusetts: Boston College Department of Economics; 2003.
40. Abadie A, Drukker D, Herr JL, Imbens GW. Implementing matching estimators for average treatment effects in Stata. *Stata J*. 2004;4:290-311.
41. Becker SO, Ichino A. Estimation of average treatment effects based on propensity scores. *Stata J*. 2002;2:358-377.
42. Ho D, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. Journal of Statistical Software. In press.
43. Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: The matching package for R. Journal of Statistical Software. In press.
44. Hansen B, Klopfer SO. Optimal full matching and related designs via network flows. *J Comput Graphical Stat*. 2006; 15:609-627.
45. Shadish WR, Cook TD. Design rules: more steps towards a complete theory of quasi-experimentation. *Stat Sci*. 1999; 14:294-300.