Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 18, Number 13, November 2013 ISSN 1531-7714

Comparing Propensity Score Methods in Balancing Covariates and Recovering Impact in Small Sample Educational Program Evaluations

Clement A. Stone & Yun Tang
University of Pittsburgh

Propensity score applications are often used to evaluate educational program impact. However, various options are available to estimate both propensity scores and construct comparison groups. This study used a student achievement dataset with commonly available covariates to compare different propensity scoring estimation methods (logistic regression, boosted regression, and Bayesian logistic regression) in combination with different methods for constructing comparison groups (nearest-neighbor matching, optimal matching, weighting) relative to balancing pre-existing differences and recovering a simulated treatment effect in small samples. Results indicated that applied researchers evaluating program impact should first consider use of standard logistic regression methods with nearest-neighbor or optimal matching or boosted regression in combination with propensity score weighting. Advantages and disadvantages of the methods are discussed.

Experimental studies provide rigorous evidence for evaluating treatment efficacy by randomly assigning subjects to treatment groups. However, when random assignment is impractical or unethical, observational studies or quasi-experimental designs are often considered. Absent random assignment, any observed differences between groups may not be attributed unequivocally to an intervention or educational program. To help control for pre-existing differences, a matched-pairs design (Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002) is often proposed. In this type of design, each member in a treatment group is matched with a member of a non-treatment group using relevant variables or characteristics. However, matching on many variables is difficult to implement particularly when continuous variables are involved. Alternatively, propensity scoring methods can be used to implement a matched-pairs design (Rosenbaum & Rubin, 1983; 1985; Schneider, Carnoy, Kilpatrick, Schmitt, and Shavelson, 2007).

A propensity score is a single summary score that represents the relationship between multiple observed characteristics for group members and treatment group assignment. It has been described as the "propensity towards exposure to treatment…given the observed covariates" (Rosenbaum & Rubin, 1983; pg. 47). This single score considers simultaneously all the relevant characteristics and attempts to reduce selection bias by weighting the characteristics relative to their influence on predicting treatment group assignment (Rudner & Peyton, 2006). The idea underlying propensity score matching is that if a member of the treatment group is matched with a member of the control group (propensity score matching), both have the same probability of being in the treatment condition (i.e., the same assumption underlying random group assignment designs). Further, Rosenbaum and Rubin (1983) showed that, in large samples, treatment and control groups matched on a propensity score will be similar relative to the characteristics used to compute propensity scores. Thus, "…if treatment and control groups have the same distribution of propensity scores, they have the same distribution for all observed covariates, just like in a randomized experiment" (Rubin, 2001; p. 171). Note that propensity scores can also be used to reduce selection bias by using the scores to weight differentially treatment and control cases (propensity score weighting).

Researchers have discussed different methods for estimating propensity scores and different procedures for using propensity score estimates to create comparison groups, and there is on-going debate as to

the relative merits of different approaches (An, 2010). Further, research comparing different approaches has yielded different recommendations. Examples of studies involving real data comparisons and simulation studies include Wilde & Hollister (2007), Harder, Stuart, Anthony (2010), and An (2010), Luellen, Shadish, and Clark (2005), Austin (2010), and Luellen (2007). These researchers found that the choice of the method and the context of the evaluation can impact the assessment of a treatment effect. Wilde & Hollinster concluded that "further research is needed before policymakers rely on [propensity score matching] as an evaluation tool" (p.455).

While the literature contains studies examining different propensity scoring methods and assumptions underlying the use of the methods, there is a lack of evidence when treatment group sample sizes are small. Small intact samples are not uncommon in educational program evaluation since educational interventions may be time intensive and difficult to implement on a larger scale. The purpose of this study was to address this specific context and compare different propensity scoring methods in a simulation study. The simulation study compared a randomized design with a variety of different methods for estimating propensity scores in combination of methods for constructing comparison groups under conditions which applied researchers commonly encounter in studies evaluating instructional or educational programs.

Propensity Score Estimation

A propensity score is defined by Rosenbaum and Rubin (1983) as the conditional probability of being selected into the treatment group given a set of covariates or observed characteristics for group members: $p(\mathbf{X}) = Pr\{Tr = 1 \mid \mathbf{X}\} = E\{Tr \mid \mathbf{X}\}$, where $Tr = \{0, 1\}$ is an indicator variable for treatment group selection and $\mathbf{X}$ is a multidimensional vector of covariates. Propensity scores therefore describe the likelihood that a population member would have been selected into the treatment group based on a set of model covariates, given they were eligible. Propensity score estimates are then used to construct a comparison group, and the average treatment effect (τ) based on a outcome measure (Y) is then estimated as follows: $\tau = E\{Y_1 \mid Tr=1\} - E\{Y_0 \mid Tr = 0\}$.

In contrast to randomized designs, propensity scoring methods rely on a set of covariates to model the treatment group selection process, and the methods cannot adjust for relevant unobserved covariates or "hidden selection bias". Propensity scoring therefore assumes observations with the same propensity score have the same distributions for observable and unobservable characteristics independent of treatment group status. Thus, for a given propensity score, treatment and control group members should be on average identical or exchangeable. This links propensity scoring to the assumption of ignorable treatment group assignment and to the corollary that the estimate of τ is unbiased (Rubin, 1997).

In the context of social and educational program evaluation research, the treatment or intervention effect could reflect attainment of program outcomes or efficacy of an intervention. For example, a treatment effect could be evaluated using a difference in mean scores on an instrument assessing program outcomes, a mean score on a test reflecting achievement outcomes, or be based on the frequency of success or completion by individuals participating in a program (i.e., odds of a successful intervention).

A common approach for modeling the treatment selection process or estimating propensity scores is logistic regression (LR) with treatment group assignment (1=Tr, 0=C) as the dichotomous outcome and a set of measured covariates as predictors (Rosenbaum & Rubin, 1983; D'Agostino, 1998). Based on the estimated model, predicted probabilities for being assigned to the treatment group (propensity score estimates) may be obtained for both the treatment group and potential control group members. However, simulation studies have found that logistic regression methods are sensitive to the functional form of the relationship between the set of covariates and treatment selection (McCaffrey, Ridgeway, & Morral, 2004). In response to this issue, McCaffrey et. al. discussed the use of generalized boosted-regression modeling (GBM) which is a nonparametric approach that recursively partitions the data for each covariate. Each partition allows for additional interactions between variables and the algorithm selects partitions which provide the most information about the outcome, in this case treatment assignment. An advantage of this approach over LR is that a large number of covariates can be used and the correct functional forms for each covariate and interactions between covariates do not have to be specified. The reader is referred to Luellen, Shadish, and Clark (2005) for an introductory treatment of the two methods for estimating propensity scores.

Using propensity score estimates to construct comparison groups assumes that the propensity scores are known for each observation (treatment and control group members), and their use therefore does not consider possible uncertainty in the propensity score estimates. Alternatively, a Bayesian propensity score analysis (BPSA) can be used (McCandless, Gustafson, & Austin, 2009; An, 2010; Kaplan & Chen, 2010). A Bayesian approach generates a distribution of propensity score estimates for each treatment group member. Using BPSA, it is possible to sample repeatedly from these distributions, construct multiple replications of treatment and comparison groups, and estimate a treatment effect for each replication of treatment and comparison groups. A distribution of estimated treatment effects can therefore be obtained that considers the uncertainty in estimating the propensity scores. A Bayesian approach could be particularly useful when there is less stability in the estimation of propensity scores such as when there is a small treatment group relative to the control group population.

Creating Comparison Groups Using Propensity Score Estimates

Once the propensity scores are estimated for all members of the population (treatment and potential control group members), there are different methods for creating comparison groups (c.f., Luellen et. al., 2005 for an introductory treatment of the methods): 1) construct matched samples; 2) construct subgroupings or stratifications on the propensity scores; and 3) weight each treatment and control group member. For all of these methods, the distributions of propensity score estimates for the treatment and control group should overlap substantially and researchers should evaluate the extent of overlap prior to constructing comparison groups.

Propensity Score Matching (PSM). PSM reflects a class of methods frequently used to construct comparison groups. For one method, a control group member with the closest propensity score to a treatment group member is matched without replacement (nearest neighbor matching). All remaining control group members are disregarded. One disadvantage of this approach is that if a match is not found for a treatment group member, there is a loss of treatment group members. Loss of treatment group members, in turn, could produce biased treatment effect estimates and loss in power to detect a treatment effect. Also, a number of control group members typically have approximately equal propensity scores. Thus, any variation due to different control group matches for each treatment group member is not considered. Note that there are a number of variants to this type of matching algorithm (c.f., Guo & Fraser, 2010).

A more recently discussed alternative, optimal matching, may also be used. The goal of optimal matching is to find a matched set of treatment and control group members, from all possible sets of matched pairs, which minimizes the total difference between propensity scores for matched pairs. Gu and Rosenbaum (1993) found that "…optimal matching is often better than nearest neighbor matching when the goal is to minimize the average distance within pairs…" (p.413). For all PSM methods, the treatment effect can then be analyzed by comparing outcome variables for the two matched groups.

Propensity Score Stratification (PSS). This method ranks all members (treatment and control group) by propensity score estimates and creates subclasses or groups of treatment and control group members that have similar propensity scores. Typically five subclasses are formed with approximately the same number of members (Rosenbaum & Rubin, 1983). For each stratum, the average treatment effect (τ_i) is computed and a weighted combination of these treatment effects is computed to evaluate treatment impact. An advantage of this method is that all treatment and control group members factor into the evaluation of the treatment effect. However, this method works best when the members within strata are homogenous in regard to the propensity score, and strata based on the same sample size do not guarantee this condition is met. A further disadvantage is that, in a small study sample, a subclass may contain a very small number of treatment group members or only control group members.

Propensity Score Weighting (PSW). PSW uses the estimated propensity scores (PS) to weight all treatment and control group observations when estimating a treatment effect. Different types of weights may be used to estimate two different effects: 1) average treatment effect or ATE (weights are 1/PS for each treatment group member and 1/(1-PS) for each control group member) or 2) average treatment effect for the treated or ATT (weights are 1 for each treatment group member and PS/(1-PS) for each control group

member). Although ATE weights are often used in applied research, Heckman (2005) discussed that in many policy contexts the effect of interest is often the average treatment effect for the treated. In choosing between ATE and ATT sampling weights, Guo & Fraser (2010) write "…in deciding whether a policy is beneficial, our interest is not whether on average the program is beneficial for all individuals [i.e. ATE] but whether it is beneficial for those individuals who are assigned or who would assign themselves to the treatment [i.e., ATT]" (p.47).

An advantage of PSW is that all possible control group members are assigned sampling weights – propensity score estimates are used to weight the treatment and ALL control members when estimating the treatment effect. Thus, there may be increased power to detect a treatment effect. Despite its advantages, PSW has its own limitations. Simulation studies (Freedman & Berk, 2008; Kang & Schafer, 2007) have shown that the PSW is sensitive to the misspecification of the propensity score model (variables and functional forms) particularly when some estimated propensity scores are small. Also, Harder, Stuart, & Anthony (2010) discuss that very small propensity score estimates and in turn very large sampling weights can be "influential" and produce biased estimates of a treatment effect when using ATE-based sampling weights.

Methodology

Using simulation methods, this study compared the effectiveness of different propensity scoring applications in balancing the measured covariates and recovering/detecting a simulated treatment effect under conditions which applied researchers face in studies evaluating instructional or educational programs. The context of this study was small scale educational program evaluations (i.e., small treatment groups) that involve a set of predetermined or intact treatment group members and a change in achievement results for a treatment versus control group.

Rather than simulate data for a population of treatment and control group members, a dataset from a state assessment program that included achievement results and commonly available covariates were used to match a set of predetermined treatment group members with members from a population of control group members. The advantage of using a real dataset of covariates was that it allowed for determining whether a set of commonly available covariates to applied researchers can be used effectively in combination with different propensity scoring methods in educational program evaluations.

The data for the study were a set of middle school achievement results for all school districts (386 districts) from a state assessment program. The covariates used to estimate propensity scores included: *proportion of economically disadvantaged students (Prop_Disadv), proportion of minority students (Prop_Minority), proportion of IEP students (Prop_IEP), overall attendance rate (Attendance), graduation rate (Graduation), and baseline test score performance (Reading and Math scale scores: SS_Math, SS_Read).* Correlations between the covariates were generally moderate, ~ | .3 to .5 |, with the exception of correlations between the scale scores *SS_Math* and *SS_Read* (> .8) and the scale scores with *Prop_Disadv* (< -.7). Note that this dataset reflects typical data that is publicly available to researchers evaluating educational interventions, that is, typical covariates and results at an aggregated level (e.g. school or district) rather than individual student results.

Study Design Factors

Sample Size of Treatment Group. Samples of 30 and 60 were chosen to represent smaller treatment group sizes that are consistent with typical educational program evaluations.

Selection of Treatment Group Members. Three conditions were manipulated: 1) random selection of treatment and control group members ("true" experiment as baseline condition); 2) non-random or predetermined selection of treatment group members with no hidden covariate; and 3) non-random or predetermined selection of treatment group members with a hidden covariate. These latter two conditions (no-hidden vs. hidden covariate) were designed to manipulate the ignorability of treatment group assignment assumption within the treatment selection process.

To implement the non-random selection of treatment group members, treatment group members were selected from a subset of the population of statewide school districts based on the covariate, *Prop_Disadv*. Specifically, treatment group members were selected from members with a value greater than the median for the covariate *Prop_Disadv*. This condition models non-equivalent groups and considers the case where an intervention is focused on disadvantaged populations. While this is admittedly a

simplistic treatment assignment model, it does reflect a target population for many educational program evaluations. Note that the unselected members from this subpopulation became part of the pool of potential control group members along with school districts below the median for the covariate *Prop_Disadv*, so that overlap in propensity score estimates between treatment and control group members was maintained.

To implement the hidden versus no-hidden conditions, the covariate used to select treatment group members, *Prop_Disadv*, was included in the treatment group selection model for estimating propensity scores for the no-hidden covariate condition and excluded from the model for the hidden covariate condition. Note that the no-hidden covariate condition reflects an ideal condition for estimating propensity scores and therefore a condition that should meet the ignorability of treatment group assumption.

Propensity Score Estimation Method. In the current study, logistic regression (LR), generalized boosted modeling (GBM) and a Bayesian propensity score analysis using logistic regression (BLR) were compared. LR models were estimated using SAS 9.2 (SAS institute, 2008). The GBM algorithm was implemented using the R routine *twang* (McCaffrey et. al., 2004; Ridgeway, McCaffrey, Morral, Burgette, & Griffin 2012). PROC MCMC in SAS was used to estimate BLR models with non-informative priors for all coefficients.

Method for Constructing Comparison Groups. Three commonly used and/or researched methods for assigning control group members were compared: 1) Nearest neighbor or greedy matching; 2) Optimal matching; 3) Using propensity scores as sampling weights (estimating both ATT and ATE). The nearest neighbor matching was conducted using SAS macro %GREEDMTCH (Parsons, 2001). The SAS macro *Proc Assign* (Coca-Perraillon, 2007) was used for optimal matching (see Stuart's website for other program options – http://www.biostat.jhsph.edu/~estuart/propensityscoresoftware.html). For the current study, propensity score stratification was not evaluated, since the number of stratifications and sample size suggested by Rosenbaum and Rubin (1984) limits the use of this method with smaller treatment groups. Further, the matching methods involved paired matching, or matching a single treatment group member to a single control group member, rather than matching multiple control group members with a single treatment group member. This approach is one of the basic and common matching methods used, and focusing on this method served to reduce the number of experimental conditions under study.

It might be argued that ATT sampling weights are more consistent with the nature of the non-random selection of treatment group members used in the present study. As Guo and Fraser (2010) suggested, when the focus is estimating an effect for those individuals who are assigned to treatment rather than an average effect for all individuals in the population, evaluating ATT is more appropriate. However, since ATE sampling weights are commonly used both sets of weights were evaluated. Also, while in principal propensity score estimates from *twang* can be used with propensity score matching (McCaffrey, personal communication, November, 2012), the *twang* program was specifically designed to be used with propensity score weighting.

Data Generation and Analysis

For each combination of conditions described in the *Design Factors* section, the data generation and analysis steps included:

1) For one sample size condition (n=30 or 60), select a set of treatment group members without replacement from the population of districts in the state achievement dataset (386 school districts) using one of three methods (random or two nonrandom conditions). All unselected school districts form the pool of potential control group members (360 - n).

2) Simulate a random small standardized mean effect size (mean d = 0.2) as the program impact on achievement results for only the selected treatment group members (Cohen's criteria, 1988).

3) Estimate propensity scores for the treatment group members using each of the three methods (LR, GBM, and BLR). Note that estimation of propensity scores was based on all included covariates as suggested by Rubin and Thomas (2000).

4) Create a comparison group from the pool of potential control group members using the propensity score estimates from Step 3 (matching methods, sampling weights).

5) Estimate the treatment effect using analysis of covariance with baseline math scale scores as the covariate for each combination of propensity score

estimation methods and methods for constructing comparison groups from Steps 3 and 4 (propensity score estimation method crossed with methods for constructing comparison groups). Note that this covariate was also included in the model used to estimate propensity scores since adjusting for pre-treatment differences is important to the assessment of a treatment effect (Schafer & Kang, 2008). For the BLR approach, the treatment effect was evaluated by sampling from the posterior distributions for propensity scores for each treatment group member, creating different samples of treatment and comparison groups, evaluating the treatment effect in each of the samples, and using SAS PROC MIANALYZE to combine results across samples (imputations) to obtain a valid hypothesis test for the treatment effect. This approach isolates the impact of uncertainty in propensity score estimates (Kaplan & Chen, 2010).

6) Repeat the experiment (Steps 1-5) to obtain 500 replications or a distribution of balance check criterion statistics and estimated treatment effects. Recovery of the simulated treatment effect was evaluated by examining the bias and root mean squared error (RMSD) of the simulated treatment effect across replications. The balance check criterion was evaluated for each measured covariate by the standardized difference test – absolute difference in the sample means for the treatment and control groups divided by the pooled standard deviation for each measured covariate (Austin, 2007). Finally, since a treatment effect was computed for each replication, the empirical power of the test (i.e., the number of times a treatment effect was significant across replications) could be computed and compared across conditions. It should be noted that in a "real" application of propensity scoring methods, the propensity score model would be adjusted if inadequate balance in the measured covariates was attained. However, such an approach is difficult to implement in a multiple replication simulation study. As will be discussed, except for a few conditions, adequate balance in the measured covariates was attained.

7) Repeat Steps 1to 6 for each combination of sample size and selection of treatment group conditions.

Results

The results describe the degree to which the various propensity scoring applications were able to: 1) balance pre-existing differences in the measured covariates (balance check criterion); 2) recover the simulated treatment effect (mean d=.2); and 3) detect a significant treatment effect (empirical power). Results are reported separately for the two non-random treatment selection conditions (no-hidden covariate and hidden covariate conditions) across the two sample size conditions (30 and 60). Results for the randomized design condition are embedded in the tables to provide direct comparison with results from the non-random treatment selection conditions.

Balancing Pre-Existing Differences using Different Propensity Scoring Methods

Checking balance in measure covariates across treatment and control groups is important since propensity scoring assumes cases with the same propensity score have the same distributions for observable and unobservable characteristics. Table 1 presents the average of standardized differences for the covariates across replications for the non-random treatment selection condition with the treatment selection covariate (*Prop_Disadv*) included in the propensity score model (no-hidden covariate condition). Also included are differences associated with the randomized design condition as well as differences prior to propensity score adjustment.

As can be seen, small standardized differences were observed for the randomized design condition with, as expected, smaller differences as sample size increased from 30 to 60. As for pre-existing differences and the presence of non-equivalent groups, large standardized differences between the groups prior to the propensity score adjustment with the exception of the *Prop_Minority* and *Prop_IEP* covariates were found. In particular, the largest differences were observed for the *Prop_Disadv* covariate which should not be surprising since this variable was used to select treatment group members.

As for how well the different propensity score approach adjusted for pre-existing group differences in the covariates, it can be seen that the top performing method (smallest standardized differences) was GBM in combination with PSW and ATT sampling weights. It can also be seen that the matching methods (Nearest

Table 1. Covariate Balance for Non-Random Treatment Group Selection with Selection Covariate Included in the PS Model (Not-Hidden Covariate Condition)

N	Covariates	Randomized Design	Pre-PS Adjustment	LR				GBM				Bayesian LR			
				Nearest Neighbor	Optimal	ATE	ATT	Nearest Neighbor	Optimal	ATE	ATT	Nearest Neighbor	Optimal	ATE	ATT
30	P_Disadv	.21	1.03	.13	.13	.63	.21	.35	.17	.85	.11	.26	.20	.61	.40
	P_Minority	.22	.18	.16	.15	.17	.04	.28	.16	.12	.06	.17	.17	.31	.20
	P_IEP	.22	.24	.14	.14	.14	.05	.28	.14	.15	.05	.19	.17	.24	.19
	SS_Read	.22	.70	.12	.12	.44	.11	.35	.16	.53	.09	.21	.18	.47	.27
	SS_Math	.21	.70	.13	.12	.44	.12	.36	.16	.52	.09	.21	.18	.47	.28
	Attendance	.22	.46	.14	.14	.27	.06	.32	.16	.32	.06	.19	.17	.32	.22
	Graduation	.21	.30	.15	.15	.18	.07	.29	.14	.19	.06	.18	.18	.26	.24
60	P_Disadv	.15	1.09	.08	.07	.51	.47	.22	.09	.90	.09	.16	.13	.48	.58
	P_Minority	.15	.14	.11	.11	.23	.07	.18	.10	.08	.05	.13	.12	.30	.20
	P_IEP	.14	.24	.10	.09	.11	.10	.19	.09	.17	.05	.13	.12	.16	.18
	SS_Read	.15	.75	.08	.07	.38	.24	.22	.09	.59	.07	.12	.11	.36	.33
	SS_Math	.15	.74	.08	.07	.38	.27	.22	.09	.59	.07	.12	.11	.35	.36
	Attendance	.15	.50	.09	.09	.25	.12	.22	.09	.38	.05	.13	.12	.24	.21
	Graduation	.15	.32	.10	.09	.19	.14	.19	.09	.24	.05	.12	.12	.21	.25

Neighbor and Optimal) preformed similarly and as well or better than the randomized design condition with the exception of the nearest neighbor method in combination with GBM estimation of propensity scores. The poor performance of this one method was surprising but may be explained by the number of unmatched treatment group members. For the combination of GBM and nearest neighbor matching, approximately half the selected treatment group members on average were not matched to potential control group members (i.e., a match to one significant digit could not be found**).** On the other hand, when all treatment group members were matched using optimal matching and GBM propensity score estimates, results were similar to the matching methods using LR models to estimate propensity scores. As to why the GBM methods yielded so many unmatched treatment group members in comparison with LR methods, more research is required. One possibility may be related to the small sample sizes. Luellen (2007) discussed that GBM may be a useful alternative to LR for large sample sizes. Also, the finding that propensity scoring methods were able to balance pre-existing differences more than the randomized design may be due to the fact that differences among groups can still exist with randomized designs, particularly when there are smaller sample sizes.

With regard to using Bayesian LR methods to estimate propensity scores, slightly higher standardized differences were observed in comparison with standard LR methods. Although Bayesian approaches model additional sources of error, this error typically affects variances rather than means. Rather, the higher standardized differences may be due to the non-informative prior that was used, but this requires further study to explain this result.

Finally, comparing the use of ATT versus ATE sampling weights with PSW in Table 1, results based on ATT were closer to results based on matching methods whereas results based on ATE exhibited markedly higher standardized differences. This may be expected given research findings that ATT and ATE sampling weights yield different results under non-random treatment selection conditions (e.g., Harder, Stuart, & Anthony 2010; Imbens, 2004). Further, these researchers discussed that ATT weighting was more consistent with matching methods, that is, ATT weighting and matching methods consider the treatment group the standard population (ATT); whereas ATE weighting considers the entire population of treatment and control group members the standard population. Further, as discussed by Schaffer & Kang (2008), ATE weighting may underperform use of ATT weights when there are more extreme weights.

The same analyses were conducted for the non-random treatment selection condition when the selection covariate, *Prop_Disadv*, was excluded from the

propensity score model (hidden covariate condition). Although introduction of a hidden covariate condition violates the ignorability assumption, a similar pattern in the results for the no-hidden covariate condition (see Table 1) was observed with three exceptions: 1) the top performing method was use of LR propensity score estimates with PSW with ATT weights; 2) the ATE sampling weight condition exhibited smaller standardized differences than observed for the non-hidden covariate condition; and 3) slightly smaller standardized differences were observed under all crossed methods of estimating propensity scores and constructing comparison groups. Thus, for this study, absence of the treatment selection variable from the model used to estimate propensity scores did not affect their use in adjusting pre-existing differences.

Recovery of the Simulated Treatment Effect and Empirical Power

Table 2 presents the average treatment effect and root mean squared deviation (RMSD) or variability in the estimated effect sizes from the simulated effect (mean d = .2) across the 500 replications for the two non-random treatment selection conditions (no-hidden and hidden covariate). Since the balance check failed for a few conditions (ATE weights and GBM with nearest neighbor matching), it could be argued that assessing recovery of a treatment effect for these conditions was not appropriate. However, the recovery of a treatment effect was still evaluated for these conditions to determine the direction of any bias.

Recovery of the simulated treatment effect was similar across the two non-random selection conditions which may not be surprising given the similarity in adjusting for pre-existing differences in the covariates for the two non-random conditions. As for how well the different propensity score methods could be used to recover the simulated treatment effect, it can be seen that the GBM with ATT weighting method and matching methods (Nearest Neighbor and Optimal) again performed similarly to the randomized design condition with the exception of the nearest neighbor method in combination with GBM estimation of propensity scores. Given this method exhibited poor performance in adjusting for pre-existing differences, it should not be surprising that this method also underperformed relative to recovering the simulated treatment effect. As for results related to adjusting for pre-existing differences, the poor performance can be explained by the number of unmatched treatment group members.

However, unlike the prior results, the Bayesian LR approach performed slightly better than the standard LR approach in recovering the simulated treatment effect, and the approaches using sampling weights to construct comparison groups (ATE and ATT) were less divergent from one another and more similar to using matching methods. It is interesting to note that the use of ATE sampling weights tended to underestimate the simulated treatment effect whereas the ATT sampling weights tended to overestimate the simulated treatment effect when used in combination with LR estimated propensity scores. As for the variability in recovering the simulated treatment effect, not surprisingly, as sample size increased RMSD decreased.

Table 3 presents the empirical power rates or the proportion of times a statistically significant treatment effect was found across replications. As for the recovery of the simulated treatment effect, empirical power rates were similar across the two non-random selection conditions. As was found above, the matching methods (Nearest Neighbor and Optimal) also performed similarly to the randomized design condition with the exception of the nearest neighbor method in combination with boosted regression estimation of propensity scores. This result can again be explained by the number of unmatched treatment group members for this condition. However, the propensity score weighting methods (ATE and ATT) both exhibited greater empirical power over the matching methods and the randomized design. This result can be explained by recalling that propensity score weighting utilizes the entire control group population for the comparison group (386 minus the number of treatment group members), whereas matching methods use a control group sample size equivalent to the size of the treatment group (30 or 60). Thus, increase in the sample size for the control group would be expected to increase the power of a statistical test for a treatment effect. There was also a slight increase in power associated with using weights based on ATT versus ATE.

Table 2. Recovery of Simulated Treatment Effect - Average Effect Size and RMSD

N	N	Random-ized Design	LR				GBM				Bayesian LR			
			Nearest Neighbor	Optimal	ATE	ATT	Nearest Neighbor	Optimal	ATE	ATT	Nearest Neighbor	Optimal	ATE	ATT
No Hidden Covariate														
Average	30	.27	.23	.25	.17	.27	.50	.31	.16	.18	.21	.22	.21	.34
Effect Size	60	.21	.22	.21	.13	.35	.37	.22	.20	.18	.19	.19	.15	.40
RMSD	30	.21	.17	.18	.13	.15	.43	.21	.13	.12	.17	.16	.18	.27
	60	.15	.13	.13	.11	.20	.27	.13	.11	.11	.13	.13	.13	.29
Hidden Covariate														
Average	30	.27	.23	.27	.12	.25	.57	.40	.12	.14	.20	.22	.16	.31
Effect Size	60	.21	.21	.22	.08	.28	.49	.33	.11	.13	.17	.17	.11	.31
RMSD	30	.21	.16	.18	.12	.13	.47	.28	.12	.12	.16	.16	.14	.24
	60	.15	.12	.12	.13	.13	.34	.19	.12	.11	.12	.12	.12	.20

Table 3. Empirical Power to Detect Treatment Effect

N	Random-ized Design	LR				GBM				Bayesian LR			
		Nearest Neighbor	Optimal	ATE	ATT	Nearest Neighbor	Optimal	ATE	ATT	Nearest Neighbor	Optimal	ATE	ATT
No Hidden Covariate													
30	.19	.17	.24	.63	.69	.14	.24	.74	.80	.21	.22	.60	.80
60	.40	.42	.44	.59	.68	.28	.44	.76	.77	.32	.34	.52	.90
Hidden Covariate													
30	.19	.26	.26	.65	.80	.09	.18	.71	.74	.18	.22	.49	.78
60	.40	.45	.43	.69	.86	.22	.35	.68	.76	.25	.30	.38	.84

The increase in power to detect a significant effect may be viewed as an apparent advantage to using sampling weight methods over matching methods. However, it is also possible that the increase in power to detect an effect comes at the expense of an increase in Type I error, or in other words, an increase in the probability of rejecting a null hypothesis of no difference when there is no true difference. In order to evaluate this threat to validity, the simulation study was rerun but with no simulated treatment effect (d=0). For no simulated treatment effect, the empirical power rates approximated the Type I error rate (α=.05) for the randomized design and the various combinations of approaches using matching methods for constructing comparison groups. However, for combinations of approaches using sampling weight methods for constructing comparison groups, the empirical power rates varied from .25 to .35 which is considerably higher than the nominal Type I error rate (α =.05). Thus, although propensity score weighting increases the power to detect a significant effect, it may do so at the expense of increasing the probability of identifying a non-significant treatment effect as significant (false positive). As to whether false positives versus false negatives would be preferred, the context of the intervention or educational program would need to be considered.

Other covariates were also used to model a non-random treatment group selection process (Graduation Rate and Math Scale Score) in order to evaluate whether results were affected by the particular covariate that was chosen to model the selection process. For example, the simulation study was also run with treatment group members being assigned only from districts with below average graduation rates. Results for these additional non-random treatment group selection conditions were very similar to results presented above, and are therefore not included or discussed further.

Discussion and Recommendations

In educational program evaluation research, quasi-experimental designs using propensity score approaches are often used with a relatively small intact or predetermined treatment group to evaluate program impact. However, there are various options that are

available to implement these methods. The present study used a real dataset with commonly available achievement results and covariates to compare the ability of different propensity scoring approaches to balance pre-existing differences between treatment and control groups and to recover a simulated treatment effect. In addition, a randomized design was included to provide a baseline comparison with the propensity scoring applications.

Based on this study, applied researchers using a propensity scoring approach to evaluate program impact on student achievement with small intact samples should consider the following:

1. Use of standard LR methods with either nearest-neighbor or optimal matching to construct comparison groups or GBM in combination with propensity score weighting with ATT weights assuming ATT weights are consistent with the design. These types of propensity scoring applications closely approximated a randomized design in terms of adjusting for pre-existing differences (balance criteria check) and recovering the simulated treatment effect. However, standard LR methods in combination with matching algorithms may be more accessible to researchers than approaches based on data mining methods (e.g., GBM) or Bayesian-based methods (BLR). Results also indicate that the propensity scoring applications are effective even in relatively small treatment group samples. While some applications involving propensity score weighting methods performed well, performance was more unpredictable and use of weighting methods increased empirical power at the expense of a possible increase in detecting an insignificant treatment effect.

2. In order to mitigate the potential effects of "hidden selection bias" or inadequate modeling of the treatment selection process with propensity scoring applications, use a set of covariates that are interrelated, diverse, and can account for any potential hidden covariates when adjusting for pre-existing differences (Steiner, Shadish, Cook, & Clark, 2010). In the present study, a hidden covariate was introduced into the propensity scoring process, that is, a covariate (*Prop_Disadv*) was used for treatment selection but excluded from the model for estimating propensity scores. Although introduction of a hidden covariate violates the ignorability assumption, a similar pattern in results for the no-hidden covariate condition and hidden covariate condition were found. Since available covariates (e.g, *Prop_Minority*) were related to the treatment selection variable (*Prop_Disadv*) in the present study, any negative impact of excluding the treatment selection variable from the models used to estimate propensity scores was averted.

3. Consider carefully the true treatment group population, the treatment group selection process, and the type of weights that are used if propensity score weighting is used. The present study would support use of ATT weights (i.e., evaluate treatment for treated population) over the more commonly used ATE weights when treatment group members are not selected at random. In particular, use of GBM in combination with optimal matching was effective both in terms of balancing covariates and recovering a simulated treatment effect. Although the GBM algorithm implemented in *twang* was designed for PSW, use of *twang* propensity score estimates yielded similar results when combined with optimal matching or ATT weighting.

Limitations of the Study

One of the advantages of the GBM method over LR methods is that the correct functional forms for each covariate and interactions between covariates do not need to be specified. It might be argued that, in the present study, the specific non-random treatment selection condition that was modeled did not completely capitalize on this advantage. In order to obtain a more complete comparison of the GBM and LR methods for estimating propensity scores, the LR and GBM methods were rerun under a non-random treatment selection condition that modeled an interaction between two covariates: *proportion of economically disadvantaged students (P_Disadv) and baseline Math scale score performance (SS_Math).* For this non-random condition, treatment group members were randomly selected from members with a value greater than the median for the covariate *proportion of economically disadvantaged students* and from members with a value less than the median for *baseline Math scale score performance.* Thus, all selected treatment group members (school districts) had higher than average numbers of disadvantaged students and below average Math scale score performance. Note that in the LR model, only the main effect variables were included (*P_Disadv and SS_Math).* Despite the apparent advantage to modeling the covariates under the GBM approach, similar results comparing the GBM and LR applications were obtained.

While use of the boosted regression method (GBM) for estimating propensity scores and other methods using regression trees have received a good deal of attention in the literature, this method performed poorly when combined with the nearest neighbor matching method. Though this result could be explained by the large number of unmatched treatment group members, the reason for the number of unmatched treatment group members requires additional research.

Finally, inferences and findings from any simulation study are inherently limited by the design. The context of the present study was relatively small sample educational program evaluations of student achievement outcomes. Therefore any recommendations are limited to this context and more research is required to generalize findings. In particular, although the non-random treatment group assignment model reflected a common target population for educational evaluations (i.e., disadvantaged populations), it would be useful to examine the effectiveness of the different methods with a complex multivariate treatment group assignment model. In addition, while use of a real dataset from a state assessment program with a set of commonly available covariates enhances the context of the study, future studies could evaluate which types of covariates are most relevant for propensity scoring applications in studies of impact of educational programs on student achievement: school-, student-, and/or teacher-level covariates.

References

An, W. (2010). Bayesian propensity score estimators. *Sociological Methodology*, 40(1), 151-189.

Austin, P. (2007). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27, 2037–2049.

Austin, P. (2010). The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reduction) in observational studies. *Statistics in Medicine*, 29, 2137-2148.

Coca-Perraillon, M. (2007). *Local and global optimal propensity score matching*. SUGI Global Forum 2007. Paper 185-2007.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and Analytic Issues for Field Settings*. Chicago: Rand McNally.

D'Agostino, R.B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine,* 17, 2265-2281.

Freedman, D.A., & Berk, R.A. (2008). Weighting regressions by propensity scores. *Evaluation Review,* 32(4), 392-409.

Gu, X. S. & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: structures, distances, and algorithms. *Journal of Computational and Graphical Statistics,* 2(4), 405-420.

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

Harder, V.S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods,* 15(3), 234 – 249.

Heckman, J.J. (2005). The scientific model of causality. *Sociological Methodology,* 35, 1-97.

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

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22, 523-539.

Kaplan, D., & Chen, C. J. S. (2010, March). *A Bayesian perspective on methodologies for drawing causal inferences in experimental and non-experimental settings.* Paper presented at the 2010 annual research conference of the Society for Research on Educational Effectiveness, Washington, DC.

Luellen, J.K. (2007). *A comparison of propensity score estimation and adjustment methods on simulated data.* Unpublished Dissertation, University of Memphis.

Luellen, J.K., Shadish, W.R., & Clark, M.H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review*, 29(6), 530-558.

McCaffrey, D.F., Ridgeway, G., & Morral, A.R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4), 403–425.

McCandless, L.C., Gustafson, P., & Austin, P.C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine*, 28, 94-112.

Parsons, L.S. (2001). Reducing bias in a propensity score matched-pair sample using greedy matching techniques.

Proceedings of the Twenty-Sixth Annual SAS Users Group International Conference, Cary, NC: SAS Institute Inc. Retrieved from: http://www2.sas.com/proceedings/sugi26/p214-26.pdf.

Ridgeway, G., McCaffrey, D.F., Morral, A.R., Burgette, L., & Griffin, B.A. (2012). *Toolkit for weighting and analysis of nonequivalent groups: a tutorial for the twang package.* Retrieved from: http://cran.r-project.org/web/packages/twang/vignettes/twang.pdf

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies. *Biometrika,* 70, 41-45.

Rosenbaum, P.R, & Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association,* 79, 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling that incorporates the propensity score. *The American Statistician,* 39, 33-38.

Rubin, D.B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine,* 127, 757-763.

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.

Rudner, LM.., & Peyton, J. (2006). Consider Propensity Scores to Compare Treatments. *Practical Assessment Research & Evaluation*, 11(9). Available online: http://pareonline.net/getvn.asp?v=11&n=9

SAS Institute Inc. (2008). *SAS* (Version 9.2) [Computer software]. Cary, NC: Author.

Schafer, J.L. & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods,* 13*,* 279-313.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W.H., & Shavelson, R. J. 2007. *Estimating causal effects using experimental and observational designs (report from the Governing Board of the American Educational Research Association Grants Program),* Washington, DC: American Educational Research Association.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Belmont, CA: Wadsworth.

Steiner, P.M., Shadish, W.R., Cook, T.D., & Clark, M.H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15(1), 250-267.

Wilde, E.T., & Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis and Management*, 26(3), 455-477.

Acknowledgement

The authors would like to thank both the editors and anonymous reviewers for their helpful suggestions on an earlier version of the manuscript.

Citation:

Stone, Clement A. & Tang, Yun (2013). Comparing Propensity Score Methods in Balancing Covariates and Recovering Impact in Small Sample Educational Program Evaluations. *Practical Assessment, Research & Evaluation*, 18(13). Available online: http://pareonline.net/getvn.asp?v=18&n=13

Corresponding Author:

Clement A. Stone
Professor, Research Methodology Program, Department of Psychology in Education
University of Pittsburgh
CAS [at] pitt.edu