An Illustrative Example of Propensity Score Matching with Education Research

Article · January 2012

DOI: 10.5328/cter37.3.187

CITATIONS
19

READS
10,497

4 authors, including:

Forrest C. Lane
Sam Houston State University
57 PUBLICATIONS **92** CITATIONS
SEE PROFILE

Kyna Shelley
22 PUBLICATIONS **293** CITATIONS
SEE PROFILE

Robin Henson
University of North Texas
76 PUBLICATIONS **5,294** CITATIONS
SEE PROFILE

Career and Technical Education Research, 37(3), pp. 187-212
DOI: 10.5328/cter37.3.187

An Illustrative Example of Propensity Score Matching with Education Research

Forrest C. Lane
Yen M. To
Kyna Shelley
The University of Southern Mississippi
Robin K. Henson
University of North Texas

Abstract

Researchers may be interested in examining the impact of programs that prepare youth and adults for successful careers but unable to implement experimental designs with true randomization of participants. As a result, these studies can be compromised by underlying factors that impact group selection and thus lead to potentially biased results. Propensity score matching is a quasi-experimental technique supported by the U. S. Department of Education that controls for systematic group differences due to self-selection and extends causal inference into these designs. The problem is that the method remains underutilized despite increased calls in the literature for its use. The purpose of this paper is to reduce barriers to the use of this statistical method by presenting the theoretical framework and an illustrative example of propensity score matching using SPSS (Version 20.0). Heuristic data, syntax, and a sample write-up of the analysis are provided.

Keywords: Propensity Score Matching, Selection Bias, Quasi-Experimental Design

Introduction

Experimental design is historically the only approach for estimating true treatment effects and making causal inferences. This is particularly important in the field of education research given a growing expectation of increased rigor in program evaluation (Rudd & Johnson, 2008). "Currently, only well-designed and well-implemented randomized controlled trials (RCTs) are considered strong evidence" for an intervention's effectiveness (What Works Clearinghouse, 2010, p. 11). The problem is that educational research does not always lend itself to large scale experimental design and true randomization (Grunwald & Mayhew, 2008). Unfortunately, there can be many ethical or cost implications resulting in an overabundance and often over-reliance on non-randomized studies throughout the field of education.

This being said, non-randomized designs do have a place in educational research. When done properly, they may better reflect the complexity of our educational environment (Shadish, Luellen, & Clark, 2005).

> Imagine an instructional program whose materials are thoroughly based on scientific research, but in which it is so difficult to implement that in practice teachers do a poor job of it, or which is so boring that students do not pay

> attention, or which provides so little or such poor professional development that teachers do not change their instructional practices. (Slavin, 2002, p. 19)

Research conducted in the complexity of real-world settings may provide greater generalizability of findings in practice.

The problem with non-randomized designs is that for the same reasons they may be propitious, they can also make the interpretation of treatment effects increasingly difficult. This is because non-randomized groups may systematically differ from one another based on any number of covariates (Rosenbaum & Rubin, 1983) and can lead to a biased treatment effect when these differences in the likelihood of group assignment have not been taken into account in the research design (Grunwald & Mayhew, 2008). To illustrate this problem, Shadish, Luellen and Clark (2006) conducted a study in which participants were randomly assigned to either an experiment or quasi-experiment. Then, those in the experimental condition were randomly assigned to a math or verbal training where as those in the quasi-experimental condition were given the opportunity to self-select into these same training sessions. Findings from this study suggest participants able to self-select into a specific training performed better relative to those randomly assigned to the same training in the experimental condition.

Several methods have been employed over the years to accommodate problems of endogeneity. For example, some studies have reported using regression based techniques (ANCOVA) or structural equation modeling (SEM) as methods for controlling for differences on post-test scores (Grunwald & Mayhew, 2008). However, SEM cannot fully control for all potential background variables nor is they intended to replace the issues associated with poor research design. Unless participants can be randomly assigned, researchers are subject to the interpretation of treatment effects confounded by non-randomization. This limits a researcher's ability to precisely report treatment effects and make causal inferences (Hong & Raudenbush, 2005).

When random assignment is not possible in a study, fields including medicine, statistics, and economics have been using propensity score matching to control for bias in a treatment effect of quasi-experimental designs (D'Agostino, 1998; Grunwald & Mayhew, 2008; Shadish, Luellen, & Clark, 2006). Propensity score matching is a mathematical approach to causal inference, grounded in the Rubin counterfactual framework (West & Thoemmes, 2010), that uses a participant's probability of group assignment to match or balance participants between groups. Balance rests on an assumption that propensity scores are free from hidden bias and that relevant covariates have been included in the model. By excluding participants who cannot be well matched, systematic error is reduced and extends causal inference into these designs (Rosenbaum & Rubin, 1983).

There are many implications for prospective users of propensity score matching. First, findings from matched samples may differ from those in which non-randomized groups were not equated. For example, Reardon, Cheadle and Robinson (2009) reported a smaller effect of Catholic schooling on math skills while Morgan (2001) reported a larger effect of private school education on math and reading achievement relative to non-matched samples. Other similar studies may be found in economics (Dehejia & Wahba, 2002), medicine (Austin, 2008; Schafer & Kang, 2008), and sociology (Morgan & Harding, 2006). As a result, propensity score matching may improve precision in result reporting. Additionally, the U.S. Department of Education (2003) supports propensity score matching as a method for evidence-based research when group equivalence can be established through the analysis.

> We believe that such well-matched studies can play a valuable role in education, as they have in medicine and other fields, in establishing 'possible' evidence of an intervention's effectiveness, and thereby generating hypotheses that merit confirmation in randomized controlled trials. (p. 5)

The problem is despite calls for more scientifically based methodology within education, propensity score matching remains greatly underutilized in the literature (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007; Slavin, 2002). For example, only one previous study was found within *Career and Technical Education Research* having used this technique (Rojewski, Lee, & Gemici, 2010). The purpose of this paper is to reduce barriers in the use of propensity score matching by presenting the theoretical framework and an illustrative example of this statistical method. Specifically, a demonstration of propensity score estimation using logistic regression and nearest neighbor matching within calipers will be presented.

Propensity Score Matching

The literature on use of propensity scores can be attributed to the seminal work of Rosenbaum and Rubin (1983). In true randomization, participants have an equal probability of being assigned to either a treatment or comparison group. As a result, groups can be compared to one another because systematic differences have been controlled through the experimental nature of the design. Conversely, quasi-experimental designs are subject to participant self-selection which introduces bias when comparing differences in the treatment effect between groups. This may threaten a study's internal validity given an unequal and unknown probability of group assignment. As a result, groups may not be comparable at baseline. Propensity score matching accounts for this problem by using regression techniques to predict group assignment from theoretically relevant covariates and then matches participants on these predicted scores (i.e. propensity score).

A propensity score (π) for an individual (i) is defined in Rosenbaum and Rubin (1983) as the conditional probability (P) of assigning a participant to a particular treatment or comparison group (T) given a set of covariates (X), expressed as,

$$\pi_i = P(T_i = 1|X_i). \quad (1)$$

Theoretically relevant pretreatment variables are used to derive probabilities of group membership which are then used to match participants in treatment and comparison groups such that both groups have equal means or likelihoods of receiving treatment. Once matched, any differences between these groups should be more reflective of the true treatment effects in the population and analogous to the interpretation of randomized designs.

In quasi-experimental designs, it is important to recognize that group selection can be influenced by any number of covariates leading to bias in the estimation of treatment effects. For example, a researcher examining the effectiveness of an after school program on student performance may find systematic differences in intrinsic motivation, prior academic performance, or parental involvement between participants across groups. Therefore, theoretically relevant covariates likely to predict group membership should be identified and included in the estimation of the propensity score. There are no limits to the number of covariates that may be used in this estimation process. However, these scores can "only be as good as the covariates that are at the disposal of the researcher" (Thoemmes & Kim, 2011, p. 94). Therefore, researchers should seek to identify covariates grounded in the literature that are

likely to influence treatment selection and thus provide a more meaningful and statistical approximation of group membership.

Once covariates have been identified, the probabilities of group membership or propensity scores are calculated for all participants. Logistic regression is the most commonly used estimation technique (Guo & Fraser, 2010; Thoemmes & Kim, 2011) and is relatively easy to interpret given that the predicted probabilities (P) of group membership (T) are the propensity scores (π) for a given set of covariates (X).

$$\pi_i\,(X_i) = P(T_i = 1|X_i) = \frac{1}{1+e^{-x_i b_i}} \quad (2)$$

Propensity scores may also be estimated through other methods such as bagging or boosted regression trees (Austin, 2008; Shadish et al., 2006).

Once propensity scores have been estimated for all individuals, a conditioning strategy is used to produce groups with similar means and distributions of propensity scores (Rosenbaum & Rubin, 1984). Three primary methods exist for obtaining statistically equal likelihoods of group assignment. These methods include matching, regression adjustment, and stratification (D'Agostino, 1998). The matching method controls for covariates by pairing participants across groups. This may be accomplished by either (a) matching a participant on the nearest possible propensity score, (b) matching within a caliper, (c) Mahalanobis metric matching, or (d) Mahalanobis metric matching on a specified caliper based on the average of the variances within the group. Alternatively, researchers may use an adjustment in the regression analysis, achieved by either subtracting the effect of covariates from the treatment effect or by adding the propensity score as a variable in the regression equation when estimating treatment effects. Lastly, stratification (also called sub-classification) may be used to place participants into sub-populations (groups / strata) so that participants can be compared based on the groups or strata they are assigned.

Once a conditioning strategy has been employed, balance in the newly matched sample should be examined to evaluate the effectiveness of the conditioning strategy. One method is to test each covariate in a 2 x 5 (Group x Strata) two-way ANOVA. Treatment and comparison groups are stratified across quintiles on the propensity score to assess the statistical significance of group differences and interaction effects (Rosenbaum & Rubin, 1984). Both treatment and comparison groups are assumed to be balanced when *F*-values are small with no statistically significant interaction effects (Rubin, 2002). Effect sizes should also be used given the sensitivity of these tests to sample size (Thoemmes & Kim, 2011). This may be accomplished by examining standardized mean differences (Cohen's d) between groups on covariates or the shared variance (e.g., η^2 or ω^2) between individual differences in the propensity scores and the group by strata variables. Lastly, covariate box plots or Q-Q plots of propensity scores and covariates may be used as graphical approaches to evaluate balance. Distributions of propensity scores for both groups should be similar and the plot of individual propensity and covariate scores should lie on the regression line. Propensity score matching assumes that, once balanced (i.e., statistically equal group means on propensity scores and covariates), there are no systematic differences between groups. Therefore, treatment effects can be estimated on the outcome variable(s) by testing in newly matched sample through a t-test or appropriate multi-group equivalent analysis.

Heuristic Example

Hypothetical Scenario

Reading proficiency is an important skill that must be successfully demonstrated among high school students for graduation. However, according to a government report approximately a quarter of these students cannot read at basic levels across secondary institutions in the United States (Loomis & Bourque, 2001). In an effort to address this problem, researchers have examined various instructional strategies impacting reading comprehension and motivation. One such approach includes the content area reading strategies program (CARS) and suggests that a "teacher influences, or can influence, the reading activity by teaching CARS to readers and by encouraging students' use of CARS within classroom reading" (Park & Osborne, 2007).

Given this literature, a previous study examined the impact of CARS instruction using a pre-test post-test nonequivalent-control-group design (Park & Osborne, 2007). "Students were taught three animal science lessons from the state approved curriculum and included anatomy and physiology, nutrition, and reproduction. The lessons were taught over the course of 23 school days, or nearly 1600 minutes of instruction" (Park & Osborne, 2007, p. 57). Students who received CARS instruction were then examined for differences in student comprehension of agricultural science content as well as their motivation to read these educationally related materials. Results from this study published in *Career and Technical Education Research* suggested that student receiving CARS instruction increased the number of books read per month, time spent reading for school, and showed statistical increases in the time spent on pleasure reading relative to those who did not receive this instruction.

Although these results suggested improvement in reading proficiency as a result of CARS instruction, Park and Osborne (2007) also suggested student pre-test scores, grade level, grade point average, gender, ethnicity, and standardized reading levels were statistically significant predictors of agricultural posttest scores (R^2 = .67). As such, some may be inclined to use these predictors in an ANCOVA to control for group differences as a result of the non-randomized design. The problem with this approach is that ANCOVA is inappropriate when differences between groups on covariates are large or when there is the presence of an interaction effect (i.e., heterogeneity of regression slopes) between the covariate and the treatment variable (Hinkle, Wiersma, & Jurs, 2003). Standardized mean differences between CARS participants and non-participants in this study varied between d = .35 and d = .58. As such, ANCOVA would be an inappropriate technique for examining treatment effects. The following example illustrates how propensity score matching may be used to control for non-random assignment and self-selection bias prior to the estimation of treatment effects.

Sample

A small heuristic sample (N = 30) was created using the same variables (Table 1) illustrated in Park and Osborne (2007) including grade level (9 – 12 grade), socioeconomic status, Florida Comprehensive Assessment Test (FCAT) scores, and grade point average (GPA). Efforts were also made to replicate reasonable response patterns to variables while still maintaining the ability to clearly illustrate the analysis. For example, one third of the sample was specified as qualifying for free or reduced lunches (33.3%). Approximately one quarter (23.3%) of the sample was specified as non-white with slightly more boys (n = 17) than girls (n = 13) in the study. Fifty-six percent (n = 17) of the sample scored 2 or below on the FCAT reading level with the remaining students equally distributed across the all other reading levels.

Lastly, student GPA was recorded with mean of 3.32 and a standard deviation of 0.34 grade points. Although multiple dependent variables were explored in Park and Osborne, only the number of books read per month was explored in this heuristic example to improve transparency and clarity of propensity score matching methods.

Estimation of Treatment Effects without Propensity Score Matching

Initial examination of the data suggested those receiving CARS instruction increased their reading by about half a book more per month (M = .64, SD = .84) than those who did not receive CARS instruction (M = .06, SD = .57) and this difference was statistically significant (t[28] = 2.231 p = .034). Furthermore, the magnitude of this difference was about three quarters of a standard deviation (d = .805). Although the practical importance of these findings should always be taken in context of prior research (Thompson, 2002), CARS instruction was assumed to have had a meaningful impact on student reading comprehension given the consistency of this study's findings with Park and Osborne (2007).

The problem with this interpretation is that these effects are potentially impacted by self-selection and non-randomization. Although a comparison group is used in the study (i.e., quasi-experimental design), differences between these groups may be a result of systematic differences that can bias the interpretation of this finding. For example, fewer students receiving CARS instruction qualified for free or reduced lunches and seem to have higher levels of GPA (0 = 3.22; 1 = 3.43). This may suggest that students receiving CARS instruction came into the program with a distinct advantage over those not receiving this instructional strategy. Rather than discuss this as a limitation to the study, propensity score matching could be used to better equate these two groups so that any comparisons made reflect an equal likelihood of receiving CARS instruction.

Statistical Software for Propensity Score Matching

A variety of statistical software packages are available to conduct propensity score matching. For example, the PSMATCH2 algorithm is available in Stata (Leuven & Sianesi, 2004), the SUGI 214-26 "GREEDY" Macro in SAS (D'Agostino, 1998), and the Matchit (Ho, Imai, King, & Stuart, 2007) or Matching (Sekhon, 2011) packages in R. However, SPSS is illustrated in this example as it tends to be a familiar statistical program among social science researchers. Point and click methods may be used to estimate propensity scores but the use of syntax is still required for matching on these estimated scores.

Two versions of SPSS syntax are freely available to help researchers with the matching process. The first syntax option, written by Painter (2009), conducts nearest-neighbor matching and may be downloaded from Jordan Institute for Families at the University of North Carolina at Chapel Hill. The second option utilizes the SPSS R plug-in so that functionality from the Matchit program in R can be incorporated in SPSS (Thoemmes, 2012). Given the widespread use of Matchit and greater flexibility in matching options (i.e., one-to-one, one-to-many matching) provided in Thoemmes (2012), this specific syntax was illustrated in the analysis. Readers are directed to the appendix in Theommes (2012) for instructions on downloading the SPSS R plug-in, custom dialog "PS Matching" and supplemental documentation of the program.

Covariate Selection and Estimation of Propensity Scores

The first step in the estimation of propensity scores is to select covariates likely to impact group selection (Tables 1 and 2). In other words, covariates that may help predict a student's likelihood of receiving CARS instruction should be included in the estimation process. These variables may be theoretically or empirically related through prior research. In this example, six variables were found to have explained 67% of the variability in post-test scores in the literature (Park & Osborne, 2007). As such, pre-test scores, grade level, grade point average, gender, ethnicity, and standardized reading levels were included in a logistic regression model to estimate the probability of receiving CARS instruction. A logistic regression was performed, saving predicted probabilities and logit transformations given they are used in the matching process. The dependent variable was specified to be a student's instructional group (0 = comparison group; 1 = CARS instruction group) given this is the outcome when estimating propensity scores.

Table 1
Summary of codes and variable descriptions

Variables	Variable Descriptions/Codes
id	Participant ID
CARSInst	0 = No CARS; 1= CARS
SES	0 = No Free Lunch; 1 = Free Lunch
Min	0 = White; 1 = Non-White
Gr	1 = 9^{th} Grade; 2 = 10^{th} Grade; 3 = 11^{th} Grade; 4 = 12^{th}
FCAT	1 = Little success; 5 = Highest success
Gen	0 = Female; 1 = Male
GPA	Grade point average on a 4.0 scale
Pre	Number of books read per month prior to instructional
Post	Number of books read per month prior to instructional
Δ	Change in the # of books read as a result of the
π	Propensity score or likelihood of being assigned to CARS

The logistic regression model was then examined to assess the quality of propensity scores. First, the correct classification of participants to groups when compared to the null hit rate indicated a 48.9% improvement (Table 2). An inferential goodness-of-fit test (Hosmer–Lemeshow) was also performed and suggested good model fit $\chi^2(8) = 7.664$ ($p = .467$). This seemed to imply CARS selection was not random and could be reasonably predicted as a result of the covariates identified from the literature used in the estimation of propensity scores. However, readers are cautioned that good covariate selection is only part of the process and not the sole aim of the method (Caliendo & Kopeinig, 2008). Rather, covariate selection should lead to quality propensity scores which then enable matching and provide balance on covariates.

Table 2
Logistic Regression Analysis of Participant Likelihood of Receiving CARS Instruction

Predictor	β	*SE* β	Wald's χ^2	*df*	*p*	e^{β}
SES	-1.93	1.38	1.96	1	.16	0.15
Min	-0.77	1.47	0.27	1	.60	0.46
Gr	0.38	0.38	1.02	1	.31	1.46
FCAT	0.23	0.41	0.30	1	.58	1.25
Gen	-0.17	1.15	0.02	1	.88	0.85
GPA	2.47	1.69	2.16	1	.14	11.86
Pre	-0.69	0.53	1.71	1	.19	0.50
Overall Model Evaluation			χ^2	*df*	*p*	
Goodness of Fit			8.87	8	.35	
Null Hit Rate			53.30			
Model Hit Rate			73.30			

Differences in the likelihood of receiving CARS instruction between the groups (i.e., bias as a result of self-selection) were then compared using an independent samples *t*-test. This initial assessment is important for evaluating the magnitude of bias and any improvement after propensity score matching. Results indicated the groups were statistically different in their propensity scores or initial likelihood of being selected into the CARS instructional group ($t[28] = 3.411$, $p = .002$, $d = 1.28$). Students receiving CARS instruction were nearly twice as likely to be in this group compared to those who were not when examining group means of propensity scores (Table 3). This difference was approximately one standard deviation and suggested that the two groups should not be directly compared when estimating treatment effects (Figure 1).

Nearest Neighbor Matching Within a Specified Caliper

Propensity scores were then used to match participants receiving CARS instruction to those who did not on their likelihood (i.e., propensity score) of participating in this instructional strategy. Matching on propensity scores is automated in the syntax through the use of a "greedy matching algorithm that sorts the observations in the treatment group by their estimated propensity score and matches each unit sequentially to a unit in the control group that has the closest propensity score" (Thoemmes, 2012, p. 7). However, a distance measure (i.e. caliper) was also used to inform which pairs were well matched. This distance was specified *a priori* as the standardized mean differences in the logit transformations of the propensity scores. Specifically, a caliper of 0.25 standard deviations ($d = 0.25$) of this score was used since this has been suggested as a reasonable distance for reducing bias between groups (Stuart, 2010). As a result, seven well-matched pairs of CARS and Non-CARS participants ($n = 14$) were identified in the dataset.

Post-matching Analyses to Evaluate Balance

A series of strategies were then employed to evaluate balance as a result of the propensity score matching model. First, Rubin (2001) suggests the standardized difference in the mean propensity score between the two groups should be near zero ($d < 0.20$). In this heuristic

example, the standardized mean difference was reduced from an initial group separation of d = 1.28 to a post matching group separation of d = .05 (Table 3). This represented a 96% reduction in to the initial group separation in propensity scores which were now statistically non-significant (t[12] = 0.093, p = .927). Additionally, Rubin (2001) suggests the ratio of the propensity score variances in both groups should be near one. This ratio was relatively unchanged prior to matching (0.91). However, the post-matching variance ratio was improved to 0.96 and within recommended limits (0.80 – 1.20). As a result, both group mean likelihoods for treatment and shape of those likelihood distributions were similar between groups (Figure 1).

Table 3
Covariate Balance Pre- and Post-Matching on Covariates

	Non-CARS		*CARS*					
Interval Covariates	*M*	*SD*	*M*	*SD*	*t*	*df*	*p*	*d*
Pre-Matching								
GPA	3.22	0.32	3.43	.34	1.80	28	.08	0.68
π	.33	.22	.62	.24	2.99	28	.01	1.12
Post Matching								
GPA	3.14	0.31	3.29	.29	0.91	12	.38	0.52
π	.44	.24	.46	.25	0.93	12	.93	0.05
Nominal Covariates								
Pre-Matching					χ^2	*df*	*p*	
SES					4.286	1	.04	
Min					0.403	1	.53	
GR					5.816	3	.12	
FCAT					1.832	4	.78	
Gen					2.039	1	.15	
Post Matching								
SES					0.311	1	.58	
Min*					n/a	1	n/a	
GR					3.800	3	.28	
FCAT					4.200	4	.380	
Gen					0.311	1	.577	

*As a result of matching, only one minority case was selected for both the control and treatment groups.

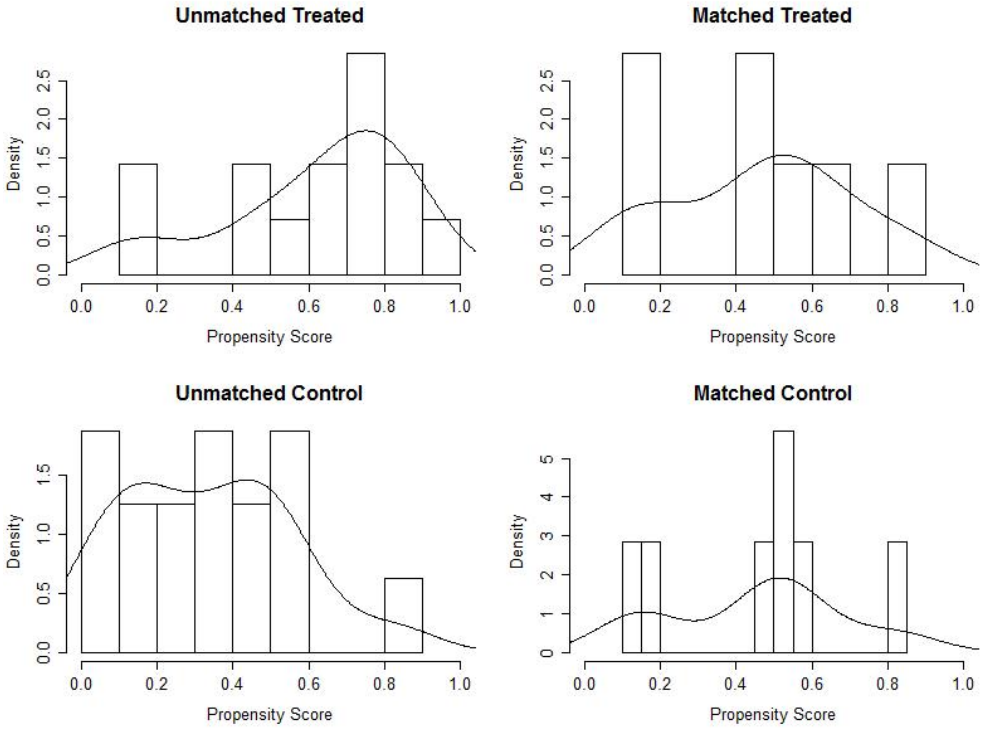

Figure 1. Distributions for Treatment and Comparison Groups of the Probability of Receiving Cars Instruction (Propensity Scores) Pre and Post Matching

Balance was also examined among covariates using a combination of univariate and multivariate approaches. Independent sample t-tests were used to examine continuous covariates (Table 3). As a result of propensity score matching, GPA between the groups was now statistically similar (t[12] = 0.905, p = .383). However, the standardized mean difference (d = 0.522) was still above recommended levels (d < 0.20). In contrast, statistical differences between observed and expected frequencies in socio-economic (SES) status was adequately mitigated in the matched sample (χ^2[1] = 0.311, p = .58).

Two multivariate tests automatically generated by the SPSS were also considered to better inform the evaluation of covariate balance after matching. The first measure, developed by Hansen and Bowers (2008), provides a test of overall covariate imbalance. It is defined as

$$d^2(z; x_1, \ldots, x_j) \coloneqq \left[d(z, x_1), \ldots, d(z, x_j)\right] \times \left(Cov \begin{bmatrix} d(Z, x_1) \\ d(Z, x_j) \end{bmatrix}\right) \times \begin{bmatrix} d(z, x_1) \\ d(z, x_j) \end{bmatrix}) \qquad (3)$$

where d is a group difference (mean difference for continuous variables) on variables denoted x, based on groups denoted by z. This statistic is analogous to Hotelling's T^2 statistic and "assesses simultaneously whether any variable or linear combination of variables was significantly unbalanced after matching" using a χ^2 distribution (Thoemmes, 2012, p. 9). Results suggested covariate balance given the non-statistically significant test result (χ^2[6] = 3.18, p = .785).

This was cross-referenced with a second multivariate imbalance measure $\mathcal{L}$, also provided in the output, which assesses the balance of all covariates including interaction effects (Iacus, King, & Porro, 2011). The $\mathcal{L}$ statistic is defined as

$$\mathcal{L}_1 = \frac{1}{2}\sum \ell_1 \dots \ell_j | t\ell_1 \dots \ell_k - C\ell_1 \dots \ell_k | \quad (4)$$

where ℓ is the frequency of a given cell, indexed by 1 to *k*, in the multivariate contingency table, for either the treatment or control group. This measure bounded by 0 and 1 and should be smaller in the matched sample relative to the unmatched sample (Thoemmes, 2012). In this example, that value was reduced from an initial value of $\mathcal{L}$ = 1.00 to $\mathcal{L}$ = .857, suggesting the overall covariate balance had been improved. Given the results of both univariate and multivariate tests, covariate balance was assumed within the newly matched dataset. However, covariate balance is not always achieved despite balance on the propensity score. Under these circumstances, the propensity score model must be respecified through "the addition of higher-order terms or interactions" (Caliendo and Kopeinig, 2008, p. 43). An alternative consideration may be to include additional in propensity score estimation matching model. The matching process would then be repeated until a new matched sample resulting in adequate bias reduction could be obtained.

Estimating Treatment Effects on the Matched Sample

Once balance is achieved on the propensity scores and covariates, groups can then be directly compared on the outcome of interest. Any differences found within this matched sample should be more reflective of the true treatment effect and analogous to experimental design. Treatment effects may be examined as either the average treatment effect on the treated (ATT), average treatment effect on the controlled (ATC), or average treatment effect on the population (ATE). Since this example compares the difference between the ATT and ATC, an independent samples *t*-test was conducted on the difference in the number of books read per month during the course of the CARS program. Results suggested this difference was smaller than initially believed to be (Table 4). Prior to propensity score matching, students participating the CARS program increased the number of books read per month by about half a book (*M* = .64, *SD* = .84) relative to non-participants (*M* = .06, *SD* = .57). This same comparison post-matching suggested CARS participants did make gains in the number of books read per month (*M* = .43, *SD* = .98) but that this difference was no longer statistically significant and resulted in a smaller standardized mean difference or effect size (*t*[12] = 0.632, *p* = .539, *d* = .338).

Table 4

Comparison of the Effect from CARS Using Matched and Non-Matched Samples

Variable	*Non-CARS*		*CARS*					
Pre-Matching	*M*	*SD*	*M*	*SD*	*t*	*df*	*p*	*d*
Reading Δ	0.06	0.57	0.64	0.84	2.23	28	.034	.805
Post Matching								
Reading Δ	0.14	0.69	0.43	0.98	0.63	12	.539	.338

Common Support Region

One of the unique features of propensity score matching is that once matched, the range in these scores provides the researcher with information about the common support region. This is defined as the shared area under the distribution of propensity scores between participants. When no support region exists, participants cannot be matched across groups. However, sufficient overlap helps to inform the generalizability of treatment effects. In this example, the common support region ranged from a propensity score of $\pi = .11$ (Participant 11) to a propensity score of $\pi = .84$ (Participant 20). As such, the result of this intervention would only be generalizable to those who were between $.84 > \pi > .11$ in terms of their likelihood of receiving CARS instruction. Participants outside this range could not be matched and contribute to bias in the estimation of treatment effects.

Discussion

Propensity score matching is a statistical technique that may be used in quasi-experimental designs to help reduce potential bias in reported treatment effects. Many studies across disciplines can be found from the literature that incorporate this technique (Dehejia & Wahba, 2002; Grunwald & Mayhew, 2008; Morgan, 2001; Morgan & Harding, 2006; Schafer & Kang, 2008; Schneider et al., 2007). However, propensity score matching remains greatly underutilized in the education research literature (Slavin, 2002). Given that the U.S. Department of Education (2003) supports propensity score matching for grant funded research, the central aim of this paper was to illustrate this technique so that interested researchers in the field of career and technical education may find greater transparency in the steps for conducting this analysis.

As with many statistical analyses, propensity score matching relies on good researcher judgment. Unfortunately, it is beyond the scope of this paper to discuss all possible variations to propensity score matching and their implications for interpreting statistical results. However, several resources can be found in the literature to help guide those looking to implement this analysis. For example, Caliendo and Kopeinig (2008) and Stuart (2010) provide a thorough discussion on the implementation of different matching methods while Thoemmes and Kim (2011) present a systematic review of the various strategies employed by social science researchers. Lastly, interested readers may also want to explore the work of Guo and Fraser (2010) which is the only known text dedicated to propensity score matching at the time this manuscript was prepared. In addition to these references, some practical considerations for propensity score matching are presented below.

First, one of the realities of propensity score matching is that some participants will be discarded from the analysis as a result of poor matching or statistically unequal probabilities of group assignment. This means that studies will likely find a reduction to statistical power in their analysis due to the reduction in sample size. Given the range of possible research questions in the literature and variables used to answer those questions, it would be difficult to provide any clear *a priori* guidance on this issue. However, larger samples are generally needed for propensity score matching (Luellen, Shadish, & Clark, 2005; Yanovitzky, Zanutto, & Hornik, 2005). "Exactly how large of a sample is needed is not clear and needs further study" (Luellen et al., 2005, p. 548). However, some have suggested that samples sizes of less than 300 may be too small for matching when prediction of group assignment is high (Lane, 2011).

Additionally, other methods beyond logistic regression are available when estimating propensity scores. Most studies (77.9%) report using logistic regression (Thoemmes & Kim, 2011, p. 103). However, other methods including classification trees, bagging, and boosted regression trees have all been suggested as possible alternative approaches (Austin, 2008; Shadish et al., 2006). Each of these estimation methods were created to help better inform covariate selection, particularly as treatment effects can be sensitive to model specification. However, there is still no consensus in the literature as to which approaches are best or the impact they would have on any substantive interpretation of treatment effects as a result of the technique employed.

Matching strategies also seem to vary greatly in the literature and should be critically examined prior to conducting propensity score matching. Thoemmes and Kim (2011) conducted a review of 86 studies using propensity score matching through 2009 in an effort to guide best practices. Most studies (43.1%) used the same one-to-one matching demonstrated in this example. However, other strategies exist such as one-to-many matching that may provide an opportunity to retain more participants. This can be beneficial especially when sample sizes of treated and untreated participants are very different, thus improving statistical power and perhaps generalizability of treatment results.

Lastly, propensity score matching relies on an assumption of strongly ignorable treatment assignment. This strong assumption suggests all relevant covariates have been included and that there are no hidden confounders (i.e., hidden bias) that could threaten the interpretation of treatment effects (Guo & Fraser, 2010). Only when this assumption is met does the method produce approximate unbiased estimates of a treatment effect (Yanovitzky et al., 2005). Sensitivity analysis may be used as a post-hoc strategy to statistically examine the impact of various levels of hidden bias on the interpretation of a treatment effect (Rosenbaum, 2010). However, sensitivity analysis is currently unavailable in SPSS and would require either the use of other statistical programs (e.g., R, Stata) or the development of additional syntax in SPSS. Sensitivity analysis may be conducted using the rbounds package in R (Keele, 2010) or the sensatt program in Stata (Nannicini, 2007; Ichino, Mealli, & Nannicini, 2008).

Conclusion

Education research rarely lends itself to large scale experimental research and true randomization. However, quasi-experimental studies can be prone to misinterpretation of treatment effects due to pre-group differences. Propensity score matching allows researchers to balance non-equivalent groups though covariates represented as a singular scalar variable. This methodology has been shown to greatly reduce effect size bias and gives non-randomized studies experimental design characteristics (Austin, 2008; Dehejia & Wahba, 2002; Grunwald & Mayhew, 2008; Luellen et al., 2005; Schafer & Kang, 2008). The following study provided an example of how propensity score matching can be implemented into non-randomized designs to minimize self-selection bias. Bias in the likelihood of group assignment (i.e., propensity score) was reduced by as much a 96% in the present example, illustrating the robustness of this technique. As matching programs like the one provided become more easily accessible on a variety of platforms, researchers should be encouraged to implement this methodology to meet the demands of a growing assessment-based climate.

References

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine, 27*, 2037-2049. doi:10.1002/sim.3150

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22,* 31-72. doi:10.1111/j.1467-6419.2007.00527.x

D'Agostino, R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of treatment to a non-randomized control group. *Statistics in Medicine, 17*, 2265-2281. doi:10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics, 84*, 151-161. doi:10.1162/003465302317331982

Grunwald, H. E., & Mayhew, M. J. (2008). Using propensity scores for estimating causal effects: A study in the development of moral reasoning. *Research in Higher Education, 49*, 758-775. doi:10.1007/s11162-008-9103-x

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

Hansen, B., & Bowers., J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science, 23*, 219-236. doi:10.1214/08-STS254

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Geneva, IL: Houghton Mifflin.

Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*, 199-236. doi:10.1093/pan/mpl013

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis, 27*, 205-224. doi:10.3102/01623737027003205

Iacus, S. M., King, G., & Porro, G. (2011). Causal inference without balance checking: Coarsened exacted matching. *Political Analysis, 20*, 1-24.

Ichino A., Mealli F., Nannicini T. (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *Journal of Applied Econometrics, 23*, 305-327. doi:10.1002/jae.998

Keele, L. (2010). *An overview of rbounds: An R package for Rosenbaum bounds sensitivity analysis with matched data*. http://www.polisci.ohio-state.edu/faculty/lkeele/rbounds%20vignette.pdf.

Lane, F., C. (2011). *The use of effect size estimates to evaluate covariate selection, group separation, and sensitivity to hidden bias in propensity score matching* (University of North Texas). ProQuest Dissertations and Theses, 115. (1041249363).

Loomis, S. C., & Bourque, M. L. (Eds.) (2001). *National assessment of educational progress: Achievement levels 1992-1998 for reading*. Washington, DC: National Assessment Governing Board.

Leuven, E., & Sianesi, B. (2004). *PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing, Statistical Software Components S432001*. Boston College Department of Economics.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction an experimental test. *Evaluation Review, 29*(6), 530-558. doi:10.1177/0193841X05275596

Morgan, S. L. (2001). Counterfactuals, causal effect heterogeneity, and the Catholic school effect on learning. *Sociology of Education, 74*, 341–374. doi:10.2307/2673139

Morgan, S. L., & Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods and Research. 35,* 3-60. doi:10.1177/0049124106289164

Nannicini T. (2007). A simulation-based sensitivity analysis for matching estimators. *Stata Journal, 7,* 334-350. doi:

Park, T. D., & Osborne, E. (2007). Reading strategy instruction in secondary agricultural science courses: An initial perspective. *Career and Technical Education Research, 32*, 45-75. doi:10.5328/CTER32.1.45

Painter, J. (2009). *Jordan institute for families: Virtual research community*. Retrieved from http://ssw.unc.edu/VRC/Lectures/index.htm.

Reardon, S. F., Cheadle, J. E., & Robinson, J. P. (2009). The effect of Catholic schooling on math and reading development in kindergarten through fifth grade. *Journal of Research on Educational Effectiveness, 3*, 45-87. doi:10.1080/19345740802539267

Rojewski, J. W., Lee, I. H., & Gemici, S. (2010). Using propensity score matching to determine the efficacy of secondary career academies in raising educational aspirations. *Career and Technical Education Research, 35*, 3-72. doi: 10.5328/cter35.102

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55. doi:10.2307/2335942

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. doi:10.2307/2288398

Rosenbaum, P. R. (2010). *Design of observational studies*. New York: Springer.

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. doi:10.1023/A:1020363010465

Rubin, D. B. (2002). The ethics of consulting for the tobacco industry. *Statistical Methods in Medical Research, 11*(5), 373-380. doi:10.1191/0962280202sm297ra

Rudd, A., & Johnson, R. B. (2008). Lessons learned from the use of randomized and quasi-experimental field designs for the evaluation of educational programs. *Studies in Educational Evaluation, 34*(3), 180-188. doi:10.1016/j.stueduc.2008.08.002

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13*, 279-313. doi:10.1037/a0014268 doi:10.1037/a0014268

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., Luellen, J. K., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review, 29*, 530-558. doi:10.1177/0193841X0575596

Shadish, W. R., Luellen, J. K., & Clark, M. H. (2006). Propensity scores and quasi-experiments: A testimony to the practical side of Lee Sechrest. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 143–157). Washington, DC: American Psychological Association.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software, 42*, 1-52.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*, 15-21. doi:10.3102/0013189X031007015

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*, 1-21. doi:10.1214/09-STS313

Stuart, E. A., & Rubin, D. B. (2007). Best practices in quasi-experimental designs: Matching methods for causal inference. In J. Osborne (Ed.), *Best practices in quantitative social science* (pp. 155-176). Thousand Oaks, CA: Sage Publications.

Thoemmes, F., (2012). *Propensity score matching in SPSS*. Available at http://arxiv.org/ftp/arxiv/papers/1201/1201.6385.pdf.

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*, 90-118. doi:10.1080/00273171.2011.540475

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling & Development, 80*, 64-71. doi:10.1002/j.1556-6678.2002.tb00167.x

U.S. Department of Education, Institute of Educational Sciences. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, DC: Institute of Education Sciences.

What Works Clearinghouse. (2010). *Procedures and standards handbook* (version 2.0). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf

West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods, 15*, 18-37. doi: 10.1037/a0015917

Yanovitzky, T., Zanutto, E., & Hornik, R. (2005). Estimating causal effects of public health education campaigns using propensity score methodology. *Evaluation and Program Planning, 28*(2), 209-220. doi: 10.1016/j.evalprogplan.2005.01.004

The Authors

Forrest C. Lane, PhD, is an assistant professor in the College of Education and Psychology, Department of Educational Studies and Research at the University of Southern Mississippi. His

primary research interests include the use of matching methods for non-equivalent group design and the development of socially responsible leadership within higher education. Email: forrest.lane@usm.edu Phone: 601-266-4556.

Yen M. To, PhD, is an assistant professor in the College of Education and Psychology, Department of Educational Studies and Research at the University of Southern Mississippi. Her research revolves around the investigation of socio-cultural influences on academic achievement and learning. Specifically, her core interest resides in examining the influence of the school environment, cognitive capabilities/disabilities, and family variables on students' learning and achievement. Email: yen.to@usm.edu Phone: 601-266-4562.

Kyna Shelley, PhD, is Professor and Coordinator of Educational Research at The University of Southern Mississippi. Her research interests include organizational behavior in higher education and research and statistics pedagogy. Email: kyna.shelley@usm.edu. Phone: (601) 266-4578.

Robin K. Henson, PhD, is a Professor of Educational Psychology in the Department of Educational Psychology at the University of North Texas. His research interests include applied statistics and measurement, reliability generalization, and self-efficacy theory. Email: robin.henson@unt.edu. Phone: (940) 369-8385.

Appendix A

Heuristic Data for a Reading Instruction Intervention in Secondary Agricultural Science Courses (N = 30)

ID	CARS	SES	Min	Gr	FCAT	Gen	GPA	Pre	Post	Δ	π
1	0	1	0	3	4	0	2.90	4.00	4.00	0.00	.049
2	0	0	1	4	5	1	3.10	3.00	2.00	-1.00	.453
3	0	1	0	3	3	0	2.85	4.00	4.00	0.00	.035
4	0	1	0	3	5	0	2.75	2.00	3.00	1.00	.149
5	0	0	0	3	3	1	3.25	1.00	1.00	0.00	.817
6	0	0	0	1	1	0	3.45	3.00	3.00	0.00	.395
7	0	0	1	1	4	1	3.50	3.00	2.00	-1.00	.363
8	0	1	0	4	2	0	3.35	1.00	1.00	0.00	.533
9	0	1	0	2	2	0	3.60	3.00	4.00	1.00	.201
10	0	0	0	1	2	1	3.60	3.00	3.00	0.00	.500
11	0	1	0	1	2	0	3.75	1.00	1.00	0.00	.497
12	0	0	0	3	1	1	3.21	2.00	2.00	0.00	.564
13	0	1	0	3	2	0	2.75	1.00	2.00	1.00	.151
14	0	0	0	2	4	0	2.95	4.00	4.00	0.00	.214
15	0	0	1	1	3	0	3.45	3.00	3.00	0.00	.322
16	0	1	0	1	2	0	3.05	2.00	2.00	0.00	.081
17	1	0	1	2	2	1	3.85	1.00	2.00	1.00	.833
18	1	0	1	4	4	0	3.75	4.00	4.00	0.00	.662
19	1	1	0	3	4	0	3.25	3.00	2.00	-1.00	.195
20	1	0	0	4	1	1	3.33	1.00	3.00	2.00	.835
21	1	0	0	4	1	0	3.05	1.00	3.00	2.00	.750
22	1	0	0	4	2	0	3.25	2.00	3.00	1.00	.755
23	1	0	0	1	1	1	3.35	1.00	1.00	0.00	.631
24	1	0	1	4	2	1	3.85	4.00	4.00	0.00	.575
25	1	0	0	2	2	0	4.00	2.00	3.00	1.00	.902
26	1	0	0	2	1	1	2.90	1.00	2.00	1.00	.450
27	1	0	0	1	5	1	3.45	2.00	2.00	0.00	.730
28	1	1	0	1	3	0	3.10	2.00	2.00	0.00	.111
29	1	0	0	1	2	1	3.25	2.00	3.00	1.00	.456
30	1	0	1	2	5	1	3.75	2.00	3.00	1.00	.794
M							3.32	2.26	2.60	0.33	.466
SD							0.34	1.08	0.97	0.76	.268

Appendix B

SPSS Syntax for Propensity Score Matching

```
LOGISTIC REGRESSION VARIABLES CARSInst
  /METHOD=ENTER Socio Minority GradeLevel FCATLevel Gender GPA BooksPre
  /SAVE=PRED
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).

SET SEED = 1234.
SET PRINTBACK=NONE.
BEGIN PROGRAM R.
#check if libraries exists
pt1 <- library(MatchIt, logical.return=TRUE)
pt2 <- library(RItools, logical.return=TRUE)
pt3 <- library(cem, logical.return=TRUE)
if (pt1 == FALSE) {
install.packages("MatchIt", repos="http://cran.r-project.org")
}
if (pt2 == FALSE) {
install.packages("RItools", repos="http://cran.r-project.org")
}
if (pt3 == FALSE) {
install.packages("cem", repos="http://cran.r-project.org")
}
#load library
library(MatchIt)
library(RItools)
library(cem)
#define options from SPPS
histplot <-  TRUE
jitterplot <- FALSE
indbal <- FALSE
histbal <- FALSE
output <- TRUE
detailedbal <- FALSE
match <- "nn"
dotplot <- TRUE
replace <- FALSE
ratio <- 1
matchmany  <-  ratio !=1
#define variables and formula strings
ID <- "id"
covs <- "Socio Minority GradeLevel FCATLevel Gender GPA BooksPre"
covs <- gsub("\n"," ",covs)
addlcovs <- ""
```

Appendix B (continued)

```
addlcovs <- gsub("\n"," ",addlcovs)
treat <- "CARSInst"
covsC <- unlist(strsplit(covs," "))
covsR <- gsub(" ", "+", covs)
covsA <- unlist(strsplit(addlcovs," "))
covsAR <- gsub(" ", "+", addlcovs)
covsALL <- c(covsC,covsA)
vars <- c(unlist(strsplit(paste(ID, treat)," ")),covsALL)
alldta <- spssdata.GetDataFromSPSS()
f <- paste(treat,"~",covsR)
f2 <- paste(treat,"~",covsR,"+",covsAR)
dta <- alldta[vars]
covsADDL <- as.data.frame(dta[covsA])
#clear warnings from previous runs
OK  <- function() warning("No warnings in estimation or matching procedure")
OK()
#missing value checker
if (sum(is.na(dta))> 0) {
warning("Missing values in covariates.")}
logical1 <- sum(is.na(dta))== 0
#binary checker
if(length(unique(dta[treat])[,1])!=2) {
warning ("Treatment variable not binary or missing values in treatment.")}
logical2 <- length(unique(dta[treat][,1])) == 2
#coding checker
pc <-  (sort(unique(dta[treat])[,1]) == c(0,1))
if (length(pc[pc==TRUE])!=2) {
warning ("Treatment variable not coded 0,1.
Please code treatment variable so that control units
are coded 0 and treatment units are coded 1.")
}
logical3 <- length(pc[pc==TRUE])==2
#create warnings object
w <- names(warnings())
if (logical1 & logical2 & logical3 == TRUE) {
#define timestamp for output file name
stamp <- substr(date(),9,19)
stamp <- gsub(" ","",stamp)
stamp <- gsub(":","",stamp)
prefixm <- ("m_")
prefixa <- ("a_")
mfile <- paste(prefixm,stamp,sep="")
afile <- paste(prefixa,stamp,sep="")
#do the matching to create both a regular and summary object m.0 and m.1
```

Appendix B (continued)

```
#if statements to accomodate choices of user
set.seed(1234)
m.0 <- matchit(as.formula(f),
 method="nearest",distance="logit",replace=FALSE,
ratio=1,caliper=.25,discard="none",data=dta
)
set.seed(1234)
m.1 <- summary(m.0, standardize=TRUE, interactions=TRUE, addlvariables=covsADDL,
data=dta
)
#create warnings object
w <- names(warnings())
#create table with big imbalances and detailed balance tables
sum.all<-m.1$sum.all[1:4]
sum.matched<-m.1$sum.matched[1:4]
rownames.all <- rownames(sum.all)
rownames.all <- gsub("distance","propensity", rownames.all)
rownames.matched <- rownames(sum.matched)
rownames.matched <- gsub("distance","propensity", rownames.matched)
rownames(sum.all) <- rownames.all
rownames(sum.matched) <- rownames.matched
sum.temp <- sum.matched
sum.temp$absolute <- abs(sum.matched$"Std. Mean Diff")
sum.temp <- sum.temp[order (-sum.temp$absolute),]
sum.imbalance <- subset(sum.temp, absolute > .25 |absolute < -.25)
sum.imbalance$absolute <- NULL
noimbalance <- "No covariate exhibits a large imbalance (|d| > .25)."
#write out complete data in R
distances <- as.data.frame(m.0$distance)
colnames(distances)[1] <- "ps"
weights<- as.data.frame(m.0$weights)
colnames(weights)[1] <- "weights"
m.all<-cbind(alldta,distances,weights)
#write out matched data in R
m.dta <- subset(m.all,weights!=0)
#create table with global imbalance test
#first with regular set then with extended covariate set
#if no add covs test will be untouched otherwise overwritten
#try expression to suppress error messages if no add covs are specified
chibal<-xBalance(as.formula(f),data=m.dta,report="chisquare.test")
try(chibal<-xBalance(as.formula(f2),data=m.dta,report="chisquare.test"),silent=TRUE)
chibal <- chibal$overall
rownames(chibal)[1] <- "Overall"
#create text object if global imbalance test is not available
```

Appendix B (continued)

```
nohb <- "Hansen and Bowers (2010) test of global imbalance is currently only implemented for
1:1 matching without replacement."
#create table with L1 balance measure
l1all <- imbalance(as.numeric(unlist(dta[treat])), data=dta[covsALL])
l1matched <- imbalance(as.numeric(unlist(m.dta[treat])),
data=m.dta[covsALL],weights=m.dta$weights)
l1all <- l1all$L1
l1all <- l1all$L1
l1matched <- l1matched$L1
l1matched <- l1matched$L1
l1measure <- as.data.frame(cbind(l1all,l1matched))
rownames(l1measure) <- c("Multivariate imbalance measure L1")
colnames(l1measure) <- c("Before matching","After matching")
#create pivot table for SPSS viewer
spsspkg.StartProcedure("Propensity Score Matching")
table=spss.BasePivotTable("Warning","Warnings")
rowdim=BasePivotTable.Append(table,Dimension.Place.row," ")
coldim=BasePivotTable.Append(table,Dimension.Place.column," ")
row_cat1=spss.CellText.String(" ")
col_cat1=spss.CellText.String(" ")
BasePivotTable.SetCategories(table,rowdim,list(row_cat1))
BasePivotTable.SetCategories(table,coldim,list(col_cat1))
BasePivotTable.SetCellsByRow(table,row_cat1,list(spss.CellText.String(w)))
spsspivottable.Display(m.1$nn,
title="Sample Sizes",format=formatSpec.Count)
if((matchmany | replace) == FALSE) {
spsspivottable.Display(chibal,
title="Overall balance test (Hansen & Bowers, 2010)")
}
if((matchmany | replace) == TRUE) {
table=spss.BasePivotTable("Overall balance test (Hansen & Bowers, 2010)","Warnings")
rowdim=BasePivotTable.Append(table,Dimension.Place.row," ")
coldim=BasePivotTable.Append(table,Dimension.Place.column," ")
row_cat1=spss.CellText.String(" ")
col_cat1=spss.CellText.String(" ")
BasePivotTable.SetCategories(table,rowdim,list(row_cat1))
BasePivotTable.SetCategories(table,coldim,list(col_cat1))
BasePivotTable.SetCellsByRow(table,row_cat1,list(spss.CellText.String(nohb)))
}
spsspivottable.Display(l1measure,
title="Relative multivariate imbalance L1 (Iacus, King, & Porro, 2010)")
spsspivottable.Display(sum.imbalance,
title="Summary of unbalanced covariates (|d| > .25)")
if((nrow(sum.imbalance))==0) {
```

Appendix B (continued)

```
spsspivottable.Display(noimbalance, hiderowdimlabel = TRUE, hidecoldimlabel = TRUE,
title="Summary of unbalanced covariates (|d| > .25)")
}
if (detailedbal == TRUE) {
spsspivottable.Display(sum.all,
title="Detailed balance before matching")
spsspivottable.Display(sum.matched,
title="Detailed balance after matching")
}
spsspkg.EndProcedure()
if (output == TRUE) {
#write out matched datafile in SPSS
dict<-spssdictionary.GetDictionaryFromSPSS()
psspec <- c("ps","Propensity Score",0,"F8.3","scale")
weightspec  <- c("psweight","Weight for PS",0,"F8.3","scale")
dict <- data.frame(dict,psspec,weightspec)
spssdictionary.SetDictionaryToSPSS(mfile,dict)
spssdata.SetDataToSPSS(mfile,m.dta)
spssdictionary.EndDataStep()
}
if (output == FALSE) {
#write out total datafile in SPSS
dict<-spssdictionary.GetDictionaryFromSPSS()
psspec <- c("ps","Propensity Score",0,"F8.3","scale")
weightspec  <- c("psweight","Weight for PS",0,"F8.3","scale")
dict <- data.frame(dict,psspec,weightspec)
spssdictionary.SetDictionaryToSPSS(afile,dict)
spssdata.SetDataToSPSS(afile,m.all)
spssdictionary.EndDataStep()
}
#Create tempfiles to export high quality graph of desired size
t1 <- paste(tempfile(),".png",sep="")
t2 <- paste(tempfile(),".png",sep="")
t3 <- paste(tempfile(),".png",sep="")
t4 <- paste(tempfile(),".png",sep="")
t5 <- paste(tempfile(),".png",sep="")
                                    if (jitterplot == TRUE) {
png(file=t1,width = 600, height = 600)
plot(m.0, type="jitter")
dev.off()
spssRGraphics.Submit(t1)
}
                                    if (indbal == TRUE) {
png(file=t2,width = 600, height = 600)
```

Appendix B (continued)

```
plot(m.1, interactive=FALSE)
dev.off()
spssRGraphics.Submit(t2)
}
                                    if (histplot == TRUE) {
        if(identical(replace,TRUE) | (ratio!=1)) {
   pscore.treated.matched <- sample(m.dta$ps[m.dta$CARSInst==1], 10000, replace=TRUE,
prob=m.dta$weights[m.dta$CARSInst==1])
   pscore.control.matched <- sample(m.dta$ps[m.dta$CARSInst==0], 10000, replace=TRUE,
prob=m.dta$weights[m.dta$CARSInst==0])
   pscore.control.all <- m.all$ps[m.all$CARSInst==0]
   pscore.treated.all <- m.all$ps[m.all$CARSInst==1]
        } else {
   pscore.treated.matched <- m.dta$ps[m.dta$CARSInst==1]
   pscore.control.matched <- m.dta$ps[m.dta$CARSInst==0]
   pscore.control.all <- m.all$ps[m.all$CARSInst==0]
   pscore.treated.all <- m.all$ps[m.all$CARSInst==1]
       }
png(file=t3,width = 800, height = 600)
par(ps=16)
op <- par(mfcol = c(2, 2))
hist(pscore.treated.all, xlab="Propensity Score", main ="Unmatched
Treated",freq=FALSE,breaks=10,xlim=c(0,1))
lines(density(pscore.treated.all))
hist(pscore.control.all, xlab="Propensity Score", main ="Unmatched
Control",freq=FALSE,breaks=10,xlim=c(0,1))
lines(density(pscore.control.all))
hist(pscore.treated.matched, xlab="Propensity Score", main ="Matched
Treated",freq=FALSE,breaks=10,xlim=c(0,1))
lines(density(pscore.treated.matched))
hist(pscore.control.matched , xlab="Propensity Score", main ="Matched
Control",freq=FALSE,breaks=10,xlim=c(0,1))
lines(density(pscore.control.matched ))
par(op)
dev.off()
spssRGraphics.Submit(t3)
}
                                    if (histbal == TRUE) {
png(file=t4,width = 400, height = 600)
op <- par(mfrow = c(2, 1))
hist(sum.all$"Std. Mean Diff.", xlab = "Std. difference", freq = FALSE, breaks=8,
xlim = c(-max(abs(sum.all$"Std. Mean Diff."),na.rm=TRUE) - .5,max(abs(sum.all$"Std. Mean
Diff."),na.rm=TRUE) + .5),
main = "Standardized differences before matching")
```

Appendix B (continued)

```
lines(density(sum.all$"Std. Mean Diff",na.rm=TRUE))
hist(sum.matched$"Std. Mean Diff.", xlab = "Std. difference", freq = FALSE, breaks=8,
xlim = c(-max(abs(sum.all$"Std. Mean Diff."),na.rm=TRUE) - .5,max(abs(sum.all$"Std. Mean
Diff."),na.rm=TRUE) + .5),
main = "Standardized differences after matching")
lines(density(sum.matched$"Std. Mean Diff",na.rm=TRUE))
par(op)
dev.off()
spssRGraphics.Submit(t4)
}
    if (dotplot==TRUE) {
png(file=t5,width = 400, height = 700)
var.names <- rownames.matched[1:(length(covsALL)+1)]
idx2 <- c(1:length(var.names))
m.diff.prematch  <- sum.all$"Std. Mean Diff."[1:length(idx2)]
m.diff.postmatch <- sum.matched$"Std. Mean Diff."[1:length(idx2)]
dotchart(m.diff.prematch[length(idx2):1],ylab = "", xlim=c(-
max(abs(m.diff.prematch),abs(m.diff.postmatch)),
max(m.diff.prematch,m.diff.postmatch)),labels =var.names[length(idx2):1],pch = 1, cex = 1)
points(m.diff.postmatch [1:length(idx2)], idx2[length(idx2):1], pch = 19,  cex = 1, ylim
=c(length(idx2)))
segments(0,0.1, 0, length(idx2)+1, lty = 3, col = "grey")
legend("bottomright",c("before matching", "after matching"),
cex=0.9, pch=c(1,19), box.lty=0, box.lwd=1,bg="#FFFFFF")
dev.off()
spssRGraphics.Submit(t5)
}
}
if ((logical1 & logical2 & logical3) != TRUE) {
spsspkg.StartProcedure("Propensity Score Matching")
table=spss.BasePivotTable("Warning","Warnings")
rowdim=BasePivotTable.Append(table,Dimension.Place.row," ")
coldim=BasePivotTable.Append(table,Dimension.Place.column," ")
row_cat1=spss.CellText.String(" ")
col_cat1=spss.CellText.String(" ")
BasePivotTable.SetCategories(table,rowdim,list(row_cat1))
BasePivotTable.SetCategories(table,coldim,list(col_cat1))
BasePivotTable.SetCellsByRow(table,row_cat1,list(spss.CellText.String(w)))
spsspkg.EndProcedure()
}
rm(list=ls())
END PROGRAM.
SET PRINTBACK=LISTING.
```

Appendix C

Sample Write-Up for Propensity Score Matching Analysis

First, covariates capable of explaining differences in the likelihood of receiving CARS instruction were identified from the literature (Park & Osborne, 2007). These variables included grade level, grade point average, gender, ethnicity, and standardized reading levels and were used as covariates in the estimation of propensity scores obtained through logistic regression. Propensity score quality was examined using inferential goodness-of-fit tests (Hosmer–Lemeshow) which suggested good model fit $\chi^2(8) = 7.664$ ($p = .467$) and considerable improvement over the chance hit rate ($I = .489$). These findings, along with prior theory, were considered justification for retaining covariates in the propensity score estimation model.

One-to-one matching without replacement and within a specified caliper was then used match participants (n = 30) from the CARS instructional strategy and comparison groups. Specifically, an *a priori* caliper of 0.25 standard deviations ($d = .25$) of the logit transformation of the propensity score was used given its support in the literature (Rosenbaum & Rubin, 1983; Stuart, 2010; Stuart & Rubin, 2007) and resulted in seven well-matched pairs. Balance in this new dataset was examined by comparing two multivariate tests of covariate balance. The first test assessed the linear combination of variables using a χ^2 distribution and indicated covariate balance given the non-statistically significant test result ($\chi^2[6] = 3.18$, $p = .785$). The second test compared the $\mathcal{L}$ statistic (Iacus, King, & Porro, 2011) which was reduced from an initial value of $\mathcal{L} = 1.00$ to a post-matching value of $\mathcal{L} = .857$ and suggested overall covariate balance had been improved. Lastly, group mean differences on the propensity score were compared before ($d = 1.12$) and after matching ($d = 0.05$). As a result of matching, the initial group separation in propensity scores was reduced by 96% and this difference was now statistically non-significant ($t[10] = .127$, $p = .902$), suggesting the matched dataset was balanced.

Given this balance in the propensity score and covariates between groups, an independent samples *t*-test was conducted on GPA to examine the treatment effects of CARS instruction. Results suggested that those receiving CARS instruction showed gains in reading relative to the comparison group but that this treatment effect was small ($t[12] = 0.632$, $p = .539$, $d = .338$).