Liu, L., & Ripley, D. (2014). Propensity score matching in a study on technology-integrated science learning. *International Journal of Technology in Teaching and Learning, 10* (2), 88-104.

Propensity Score Matching in a Study on Technology-Integrated Science Learning

Leping Liu & Darren Ripley
University of Nevada, Reno

Propensity score matching (PSM) has been used to estimate causal effects of treatment, especially in studies where random assignment to treatment is difficult to obtain. The main purpose of this article is to provide some practical guidance for propensity score sample matching, including definitions, procedures, decisions on each step, and methods of statistical analysis. The authors also implemented PSM in the data analysis of a study that examines factors affecting middle school students' technology-integrated science learning. Procedures of PSM are demonstrated and some cautions and tips for researchers in the field of education are included as well.

Keywords: propensity score matching, treatment effect, covariate balance, logistic regression, technology integration, science learning

INTRODUCTION

Propensity score is defined as the probability of a subject to be assigned to a specific treatment, conditional on the observed covariates (Rosenbaum & Rubin, 1983, 1984, 1985). The methods of propensity score matching (PSM) were first introduced by Rosenbaum and Rubin (1983, 1984, 1985), and it has become an alternative method for estimating treatment effects when treatment assignment is not random. "The basic idea is to find, from a large group of non-participants, those individuals who are similar to the participants in all relevant pretreatment characteristics" (Caliendo & Kopeinig, 2008, p32). When matching by propensity scores, theoretically, treated and untreated subjects who have the same propensity score will have the same distribution of observed variables (Rubin, 1997; Blackford, 2009).

Propensity score sample matching has been used for research in a variety of areas. For example, Barth et al. (2007) compared the outcomes of in-home therapy and residential care on behaviorally troubled youth. Nieuwbeerta, Nagin and Blokland (2009) studied the impact of first-time imprisonment on offenders' subsequent criminal career development. Bryson (2002) examined the effect of employees' union membership on their wages. Hitt

Leping Liu is a Professor at the Department of Counseling and Educational Psychology, University of Nevada, Reno. Darren Ripley, PhD, is the Head of Mathematics Department at the Davidson Academy of Nevada. Please contact Dr. Leping Liu at liu@unr.edu

and Frei (2002) estimated the effect of online banking on customers' profitability. Moreover, PSM methods are also applied to research in the field of education. For example, Dearing, McCartney, and Taylor (2009) analyzed the effect of the quality of early child care on low-income children's math and reading achievement in middle childhood. Wyse, Keesler, and Schneider (2008) evaluated the effects of small school size on mathematics achievement. Staff, Patrick, Loken, and Maggs (2008) discovered (rather obviously) that heavy drinking adolescents experience reduced educational attainment. These studies have demonstrated the strength of using PSM to estimate the treatment effects when a formal randomized control trial (RCT) isn't possible.

In the field of using information technology in education, over decades, studies have been conducted to examine the effect of using new technology tools, new instructional methods, new designs of media applications, or new models of technology integration on learning. Researchers are also interested in testing factors that influence outcomes of technology-based teaching and learning. Often, random assignment of treatment are not ensured in studies when convenient samples are used and the researcher has no control over the assignment of subjects to experimental or control groups (Gujarati & Porta, 2009). Obviously, the inferential results based on data from such samples are not scientifically convincing, and can contain biases that lead to invalid conclusions.

In such situations, and when the treatment sample size is large enough to achieve the expected effect size, the PSM method would be a proper means for sampling. To implement PSM, a researcher will be confronted with a set of questions and decisions. The main purpose of this article is to provide some practical guidance to researchers who want to use PSM for studies in the field of using information technology in education. The following sections will focus on:

1. when to use PSM,
2. basics of PSM, including theoretical framework and procedures of implementing PSM,
3. an example of using PSM in a study that examines factors influencing middle school students' technology-integrated science learning.

WHEN DOES PROPENSITY SCORE MATCHING APPLY?

RANDOMIZED CONTROLLED TRIAL VERSUS OBSERVATIONAL STUDY

Randomized controlled trial (RCT) is a specific type of scientific experimental design wherein the researchers randomly assign subjects participating in the experiment to either a control group, which receives no treatment, or the experimental group, which receives the treatment. After completing the experiment, researchers compare the outcomes between the control group and the experimental group to measure for differences (Chalmers, et al, 1981). The differences in outcomes between the control and experimental group are referred to as *treatment effects*. Treatment effects that are generalized by the researcher to the overall population are referred to as *average treatment effects*, while treatment effects that are generalized to individuals are referred to as *average treatment effects for the treated*. In an RCT, these two values can be assumed to be the same because subjects are randomly assigned to each group, which implies that the treated group should not systematically differ from the overall population. By virtue of the randomization process, any bias between control and experimental groups should be eliminated, within the limits of sampling error (Austin, 2011).

In an *observational study*, researchers simply follow and measure the performance of an existing treatment group and an existing non-treatment group, then compare the outcomes between the two groups. Unlike an RCT, researchers have no control over the

manner in which subjects are assigned to either groups (Porta, 2008). This is the key difference between an RCT and an observational study. More importantly, observational studies can't be used in the same manner as RCTs to generalize conclusions to a population because of this lack of control over assignment, which can infuse bias into the study.

AN EXAMPLE

For Example, a researcher wishes to measure whether the performance of students who receive web-based instructions is different from those who receive the traditional lecture-based instruction. In an RCT design, participants are randomly assigned into the treatment group (with web-based instruction) and control group (with traditional lecture-based instruction), and the measured performance (e.g. testing scores) are compared. The differences between the two groups, if found, could be interpreted as the effect of treatment, based on which the conclusions are able to be generalized to the population.

In contrast, in an observational design, the researcher collects data on the measured performance directly from two existing classes, which are already taught with web-based method in one, and lecture-based method in the other one. The comparison results, significant or not, can only be used to describe the performance of these two particular classes. Again, under such circumstance, the researcher has no control over the assignment of treatment. This is where the researcher wants to consider PSM as an alternative method to estimate treatment effect.

WHEN TO USE PSM

In summary, propensity score matching is appropriate for studies aiming to examine treatment effect but the random experimental/control grouping is not possible or applicable. In these cases, propensity score matching could be one of the methods to validate the treatment effect from reducing sampling bias or creating matched control group. For example, it may apply to the following situations:

1. a study that uses data sets from a multiple resource national educational database (especially, with the recent "big data revolution" (EOP, 2012), more and more huge and comprehensive data resources are available);
2. a longitudinal study, when loosing participants from either experimental or control group (Segal, et. al., 2007); or
3. typically in an observational study, where the researcher does not randomly allocate the treatment (Rosenbaum & Rubin, 1984).

Although in different contexts of studies, the method or procedures of PSM may vary, the basics introduced in the next section are the start point to learn and perform PSM, from which more in-depth and varied methods and skills can be developed.

BASICS OF PROPENSITY SCORE MATCHING

Propensity score matching (PSM) applies a series of statistical analyses and tests. Important concepts, considerations, and decision-making criteria during the PSM procedures are described below.

THE LOGIT MODEL AND PROPENSITY SCORES

Propensity Score. A propensity score is the conditional probability that a person will be in one group (e.g. experimental or control), given a specific set of observable covariates, sometimes referred to as the covariate vector, or $e(x) = p(Z = 1|\boldsymbol{X})$ (Rosenbaum and

Rubin, 1983). In this equation, $Z = 1$ represents the subject being assigned to the treatment group, and $\boldsymbol{X}$ is the vector representing the set of *covariates* for a given subject. The propensity score balances participants in each group. What this means is subjects in each group with similar covariate values will have similar propensity scores. The propensity score can be derived from a *logistics regression* equation. The following paragraph explains the two terms, covariates and logistics regression that were mentioned here.

Covariate. A covariate is any variable that is possibly predictive of the outcome being studied (Gujarati & Porter, 2009). In the above observational study, suppose that the experimental group (web-based instruction) performed significantly lower on the test than the control group. However, when the researcher looked at the subjects, it was discovered that the experimental group had more students who were in low SES (social economic status) than that were in the control group. Students from low SES families may not have convenient computer and internet access, which would have influence on their web-based learning activities. In this case, the proportion of low SES students in each group could be a predictor of their performance on the test, *not* the treatment (method of instruction) given. This phenomenon is referred to as confounding, since the researcher doesn't know whether the treatment or the SES of the student was the reason for the difference in performance. The SES of the students, in this case, would be referred to as a *covariate*. Selected covariates (such as gender, age, years of education, online learning experiences, or SES) could be included in the logistics regression equation to calculate the propensity scores.

Logistics Regression and Propensity Scores. In a logistics regression, different from a multiple regression, the dependent variable (DV) is not a continuous, quantitative variable, but a categorical variable that may have as few as two outcomes. Logistics regression is most often used in dichotomous, or binary outcome studies (Mertler & Vannatta, 2002). For example, the DV can be the membership of experimental or control group (where 1 as in treatment group, and 0 as in the control group). The independent variables (IV's) in logistics regression can be either qualitative or categorical data (George & Mallery, 2000), and for example, they can actually be the *covariates* for the study. Logistics regression analysis will generate a logistics regression equation, the Logit Model:

$$\ln\left(\frac{\hat{Y}}{1-\hat{Y}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i$$

This is the probability model to calculate propensity scores for the PSM, where $\hat{Y}$ is the propensity score for each subject, X_1, X_2, *... and Xi* are the covariates (the independent variables), and $\beta_1, \beta_2, \ldots$ and β_i are the weight for each covariate. Theoretically, propensity scores for each subject can be generated from this logit model, and then are used to match the subjects from experimental group to subjects in control group. These procedures can be performed with statistics program R with SPSS.

It is also important to know that the logistics model isn't a unique tool for creating propensity scores, or probabilities of group membership given a set of covariates. Other techniques using *probit models*, and *semi-* and *non-parametric* techniques are deeply explored in the literature. However, the logistics (or logit) model is most often used in the initial stages of PSA (Steiner & Cook, 2011). Because the logistics equation in PSM is only used to create propensity scores, and not for the purpose of prediction, *none of the assumptions* of logistics regression need to be met.

Counterfactuals. In statistical parlance, a counterfactual is a thought experiment in which a researcher hypothesizes how a subject would have performed had they not received the treatment they received (Lewis, 1973). The use of a propensity score as a statistical tool

for sample matching in particular has entirely to do with its ability to *balance* subjects based on their covariate matrices. In an ideal situation, all unignorable covariates would be measured and the propensity score would represent a parameter for each possible covariate matrix, **X**. This implies subjects could be perfectly matched based on their respective, identical propensity scores. Under this perfect setting, treatment assignment is entirely independent, thus two perfectly matched individuals or groups of individuals could be compared and treatment effect could be determined exactly.

For example, if a member of the control group is perfectly matched with a member of the experimental group, researchers could precisely determine what would have happened had the experimental subject *not* received treatment and the control subject *had* received treatment, and the counterfactual question could be answered with exact precision. That is, theoretically, if the propensity score can be calculated exactly, and the ignorability assumption isn't violated, possible hidden bias can be removed from a study (Steiner & Cook, 2013).

Assumptions for PSM. There are three assumptions that have to be met before valid propensity score matching (PSM) can occur. The primary assumption is the absence of nonignorable covariates, known as the *no unmeasured confounders assumption*. This assumption states that all variables (covariates) that can affect either treatment effect or outcome have to be measured (Austin, 2011), and it is the most crucial step in creating a valid model for use in PSM (Brookhart, et al., 2006). While it is impossible to collect a complete set of covariates, and thus ensure this assumption is met, researchers are compelled to explore the literature prior to any research endeavor involving PSM to determine as many covariates as possible that might influence treatment effect. This primary assumption for PSM allows the researcher to completely disregard all assumptions pertaining to logistics regression. The propensity score created is an estimated statistic, which is used to match subjects, not to determine the effect of any kind. After the propensity scores are generated for subjects in both treatment and control groups, the sample matching takes place.

The second assumption is that all covariates are adequately balanced, both before and after matching. This assumption, and the analyses used to verify the assumption will be explored in greater detail in subsequent sections.

The final assumption that must be met for valid use of PSM is the "common support" of propensity scores assumption. In other words, prior to matching, a researcher must observe the distributions of the propensity scores of both the control and treatment groups. It is best done side-by-side and the following hypothetical figure utilizes side-by-side histograms as shown in Figure 1 (Love, 2008).

Notice that in Figure 1-A, there is no overlap, or "common support" (Caliendo & Kopeinig, 2008), between the propensity scores of the control and treatment groups. In this case, the researcher will need to reconsider the use of PSM, as any matching technique is going to mismatch experimental and control units because of the large distances between propensity scores. In Figure 1-C there is a large amount of overlap, however the researcher must contend with how to approach the non-overlapping tails of the distributions. The choice of matching technique will make the decision for the researcher, in terms or whether to discard experimental units, control units or both, however to simply allow the algorithm to run without understanding potential creation of bias in the results is statistically feckless and should be avoided whenever possible.

Ideally, the distributions should completely overlap, as in Figure 1-B. It should be obvious to the reader that each treatment unit will have a close match to a control unit and vice versa. As any researcher knows, this rarely happens and rigorous explanation of choices surrounding deletion and/or mismatching of data should be included in any study.

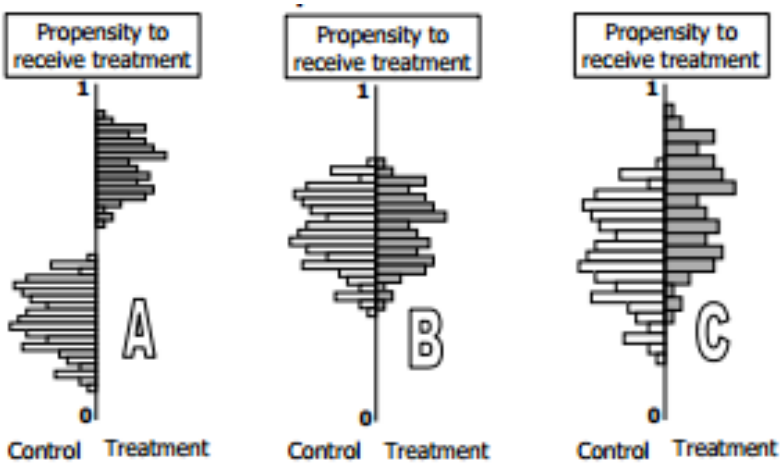

Figure 1: Distributions of the Propensity Scores of Control and Treatment Groups (Love, 2008)

DECISIONS AND PROCEDURES OF PSM

Although there are many statistical platforms with the capability to perform PSM, the learning curve is steep for researchers new to the method and often requires knowledge of programming languages. For this reason, many researchers using PSM turn to more user-friendly statistical software such as SPSS, which was the case for the study in this article. As of the writing of this article, SPSS doesn't have PSM available in its suite of options; however, there are free plug-ins such as R Essentials which are available on the internet to assist researchers in using PSM in combination with SPSS.

Three Matching Algorithms. After propensity scores have been created for all subjects using the logit function, researchers must decide which matching algorithm to use to match experimental subjects to control subjects. Among the number of different algorithms that have been invented to match subjects based on propensity score since its inception in 1983, the following three matching algorithms, among other more esoteric methods not mentioned here, have been used in PSM across many disciplines:

1. *Stratification/Subclassification*. The first and simplest posited by Rosenbaum and Rubin (1983) was stratification, or subclassification. Stratification involves the researcher breaking both groups into equal strata based on propensity score. Rosenbaum and Rubin (1984) recommend a minimum of 5 strata (quintiles) and have shown that this method removes 90% of bias in a study. The caveat with stratification is that propensity scores for both groups need to have similar distributions, i.e., if the experimental group has most of its propensity scores close to 1 and the control group has most of its scores close to 0 the researcher will arrive at biased results.

The above Figure 1-C graph displays possible distributions of propensity scores for experimental (treatment) and control groups. In this case, the top 20% of experimental units (those with a propensity score closest to 1) would be matched to the top 20% of control units and so on until all of the units are matched by quintile. Clearly, this method has the potential to create bias, as each experimental group created will systematically vary from

its control counterpart. This emphasizes the importance of "common support" after matching when using PSM.

2. *Optimal Matching*. Optimal matching was invented as a computer algorithm by Dimitri Bertsekas (1991) and seeks to minimize the overall pairwise Mahalanobis distances between experimental units and their control counterparts by using a field of research referred to as flow theory. One benefit of Optimal matching is that it creates a unique matching environment which is repeatable every time the algorithm is run

3. *Nearest Neighbor Matching*. The last matching method, the matching method used for this study, is referred to as *Nearest Neighbor* (NN). In nearest neighbor matching, the researcher/algorithm selects a treatment unit, which can either be selected at random or by simply starting with the treated unit with the highest or lowest propensity score. The control group unit with the closest propensity score, sometimes referred to as the *nearest neighbor*, is matched to the treatment unit. The researcher can choose whether to match exactly one experimental unit with one control unit (1:1) or with N control units (1:N). One limitation of Nearest Neighbor matching is the matches can depend entirely on the order in which experimental subjects are chosen, particularly when matching without replacement (Austin, 2011). A control subject can inadvertently be matched to an experimental subject that doesn't have the closest propensity score match by virtue of the order of matching, which can completely change the overall control group after matching. However, very often the appropriate matching choices are based on the test of covariate balance.

The process can also be done with or without replacement. The researcher must also choose whether to use a caliper, or propensity score interval. Noticed that when matching with replacement, overall bias decreases while variance among the covariates increases. Inversely, matching without replacement results in an increase in overall bias and a decrease in covariate variances. Furthermore, when using a caliper matching, treatment units can only be matched to control units if the control unit is within that distance/interval of the treatment unit. The bias/efficiency trade-off is identical to matching with replacement, wherein using calipers decreases overall bias but increases variance (Caliendo & Kopeinig, 2008). Researchers determine the appropriate matching choices based on the test of covariate balance; and very often different combinations of the choices are tested.

Ensuring Covariate Balance. As the entire purpose of propensity scores is to balance the covariates (Rosenbaum and Rubin, 1983), after matching, the final step is to mathematically check for covariate balance. Ensuring covariate balance is the second of two assumptions in PSM, and any imbalance of even a single covariate can potentially inject bias into a study. SPSS provides a number of algorithms and tests to do so and the following are the most often used five methods.

The Hansen and Bowers (2008) test for overall balance is the first test for covariate balance. This test is only available for analysis when matching is 1:1, and without replacement. Covariates are considered poorly balanced if the test value is significant ($p < .05$).

Relative Multivariate Imbalance, L_1, is a second test for overall balance of covariates developed by Iacus, King and Porro (2009). In the literature, there is no standard value for how big the difference of L_1needs to be before and after matching, but a reduction of L_1 will be observed when covariates are sufficiently balanced (Thommes, 2012).

Third, the summary of unbalanced covariates displays whether individual or combinations of covariates display imbalance ($|d|>.25$) after matching. SPSS provides tables of detailed balance both before and after matching to determine if covariate imbalance exists. Love (2008), among others recommends comparison of univariate

distributions of each covariate separately utilizing Cohen's d, given by $\frac{(\bar{x}_e - \bar{x}_c)}{\sqrt{S^2_{pooled}}}$, where $\bar{x}_e$ and $\bar{x}_c$ are the means of individual experimental and control covariances. This value represents the standardized differences between the means of the covariances, i.e., the imbalance for each covariance (Rubin, 2001).

It is best to measure the amount of covariate imbalance both before and after PSM. Large differences prior to matching (|d|>1) imply heterogeneity of distributions of subjects for that covariate and the researcher might consider either deleting the covariate for the study or combining the covariate with another using multivariate techniques. After PSM, imbalance is again determined by how large Cohen's d is. Stuart and Rubin (2007) recommend |d|<.25, however more conservative benchmarks of |d|<.1 have become standard (Love, 2008, Shadish et al., 2008). If Cohen's d for any individual covariate after PSM exceeds these benchmarks, again it is recommended the researcher try multivariate techniques or combining covariates to try and mitigate the imbalance. However, if covariate balance is ensured, the PSM procedure is complete and the researcher may use the two data sets as they would with any experimental and control group, with the full palate of statistical tests to determine treatment effect.

Next, dot plots provided by SPSS also graphically display whether Cohen's *d* was reduced from before to after matching. These dot plots can be used to visually determine if any covariates demonstrate imbalance both before and after matching.

Lastly, histograms that demonstrate if common support exists between the control and experimental subjects' propensity scores after matching are provided by SPSS as well. These histograms are used to ensure that the criteria of common support between matched treatment and control groups are met. A quick examination of the distributions of the treatment and control group for each matching method allows researchers to determine if PSM has created a well-balanced control and experimental group.

Summary of PSM Procedures. In the literature (Love, 2008; & Steiner & Cook, 2013), the following nine procedures are most used to implement propensity score matching:

1. Covariates are identified and data is collected.
2. Covariate data is divided into experimental and control groups.
3. Covariate matrices are input in SPSS to create a logit model. The values of the independent variable will be either a 1 if the subject participated in the treatment group or a 0 in a control group.
4. The logit model generates sample propensity scores $\hat{e}(x)$for each matrix.
5. Cohen's *d* will be used to identify any covariate imbalance.
6. In the event of serious imbalance, adjustments to the model are made.
7. A matching algorithm is chosen (e.g., in the following example, the algorithm is optimal matching).
8. After matching, Cohen's *d* again checks for serious imbalances.
9. If no serious imbalances exist, traditional statistical tests can be performed to determine whether significant differences exist between the groups.

This list provides a process template for potential users of PSM to follow in their research. While common research processes are intuitive for most readers, it is important to fully understand the ideas and concepts presented in this template by virtue of the relatively unique methods propensity score matching requires. The following is an example of using PSM for the data analysis in a study that explores middle school students' technology-based science learning.

AN EXAMPLE: USING PSM IN A STUDY TO EXAMINE TECHNOLOGY-INTEGRATED SCIENCE LEARNING

For the main purpose of this article, the current section provides an example that demonstrates the PSM and data analysis procedures, using data set from a study that examines factors affecting middle school students' technology-based science learning.

ABOUT THE STUDY AND DATA SET

Participants and Procedures. Thirty-eight middle school science teachers from rural school districts in a western state participated this study. They first completed a training and developed 22 technology-integrated science lessons including learning materials or activities at three technology integration levels: (a) using technology tools as support, (b) incorporating technology components such as digital resource/materials into lesson contents, and (c) creating learning activities to interact with technology such as collecting data for earthquake events from online seismic stations. Secondly, they developed standardized assessment rubric and system to assess student learning performance and learning outcomes.

Then, each teacher taught two lessons over a school year, one lesson per semester. Each lesson was taught to a randomly assigned experimental group and a control group. Finally, a total of 10,698 scores were collected from students in the six school district. Student scores were disaggregated by demographic variables as well.

Measurements of Learning. Student learning from these technology-integrated science lessons was measured with test scores obtained from three subscales of each lesson: (a) the *Factual Item Scale* consisting of nine 1-point questions, (b) the *Conceptual Item Scale* consisting of nine 1-point questions, and (c) the *Scenario Item Scale* consisting of one 7-point comprehensive analysis question. Cronbach's alpha that evaluates internal consistency for the *Factual Item Scale* is .86, and .83 for the *Conceptual Item Scale*. The *Analysis Item Scale* consisted of only one item, so no alpha was calculated. The overall alpha for the lesson tests was .87, suggesting that the scores were reasonably reliable for participants in the study (Green & Salkind, 2005, p. 331).

Purpose and Research Question. The main purpose of the study was to explore the factors that might have impact on technology-integrated science learning. From the overall design, such factors include (a) use of technology-integrated lesson components, (b) level of technology integration, or (c) demographic variables or group-ships. For the PSM example in the present article, the authors examined the IEP status (individualized educational plan) as the groupship factor. IEP is an education program for students who are qualified for special education services. The research question was: For each of the subscales (*Factual Item Scale*, *Conceptual Item Scale*, or *Scenario Item Scale*), are there any differences in the mean test scores between students who are in an IEP program and those who are not in an IEP program?

Before the comparison tests could be performed, PSM was conducted to identify and match the IEP group (as treatment) to none-IEP group (as control) with a random selected sample from the data set of N=10,698 subjects. The procedures are described in the following sections.

PROCEDURES OF PROPENSITY SCORE SAMPLE MATCHING

Determining the Covariates. The decision as to which covariates will be used to create the logistics regression model, and thus the propensity scores for each of the subjects, is the most important decision when using observational data to create a quasi-experimental

setting (Rubin, 2001), as the non-ignorability assumption must be met. Table 1 summarizes the independent and dependent variables used to create the logistics model: the dependent variable (*Y*) is the status of IEP, and the five covariates are gender (X_1), SES (X_2), ethnic (X_3), total test scores (X_4), and performance assessment (X_5). These five covariates are commonly used in literature to examine technology-based learning outcomes (DaDeppo, 2009; Judge & Watson, 2011; & McGraw, et. al., 2006).

Table 1. Covariates for the Logit Model

Variable	Variable Type	Coding
(X_1) Gender	Independent Qualitative/binary	1 = Male 2 = Female
(X_2) SES Socioeconomic Status	Independent Qualitative/binary	0 = Low SES 1 = Not low SES
(X_3) Ethnic Ethnic Background	Independent Qualitative	1 = Black 2 = Hispanic 3 = Asian/Pacific Islander 4 = American Indian 5 = White
(X_4) Test Subject's Total Test Score	Independent Quantitative	Score values from 4 - 64
(X_5) P_Assmt Performance Assessment Score	Independent Quantitative	Score values from 0 – 100
(Y) IEP Subject's Participation in Special Education Program	Dependent Qualitative/binary	0 = not in IEP (control group) 1 = in IEP (experimental group)

The Logit Model. Of the original 10,698 values, many of the subjects were missing data or data had been input incorrectly. Any such subjects were removed from the data set as logistics regression models cannot be created with missing data (Mertler & Vannatta, 2002), resulting in a final overall data set of size $N = 8984$ ($n_1 = 1173$ treatment and $n_2 = 7811$ control units). Because the data set was still so large, it was decided to randomly select 300 subjects from the IEP treatment group and 3600 from the non-IEP control group as the data set for the PSM. Then a logistic regression equation was generated:

$$ln\left(\frac{\hat{Y}}{1-\hat{Y}}\right) = .455 - .741x_1 + .697x_2 + .058x_3 - .118x_4 - .002x_5$$

In this equation, $\hat{Y}$ is the propensity score for each subject, X_1, X_2, *... and* X_5 are the covariates (the independent variables) selected for the PSM in this study (see Table 1), and the coefficients are the weight for each covariate. Again, none of the assumptions of logistics regression were checked or assumed to be true for this logit model, as it was only used to determine propensity scores for each of the subjects, instead of making predictions.

Choice of Matching. For the current study, the following matching methods were chosen:

- Nearest Neighbor matching algorithm
- 1 : 1 experimental-control ratio
- matching without replacement
- matching without caliper

With the combination of these choices of matching, a balanced model was expected, and was examined on the covariates balance in the next section.

Tests for Covariates Balance: Analysis and Results. A set of diagnostics tools can be used to test if the model meets the criteria for adequately balancing matched subject. For this study, the five most often used tools as described in previous section were used.

First, the *overall balance test* by Hansen and Bowers (2008) was performed, as in this study the matching method was 1:1, without replacement. The resulted test value is a chi-square (X^2) value, and when it is significant (p<.05), covariates are considered poorly balanced after matching. In this case (X^2 = .460, p = .994), there is no evidence of serious imbalance of covariates.

The second test for overall balance of covariates, *Relative Multivariate Imbalance,* L_1, (Iacus, et. al., 2009), was examined after the data was input into SPSS for analysis. Recall that there is no cutoff for whether covariates are balanced, only that a reduction of L_1 from before matching to after matching is necessary to consider a matching technique to have improved covariate balance. In this case, the decrease of L_1 from before matching (.415) to after matching (.093) does exhibit so (see Table 2).

Table 2. Relative Multivariate Imbalance, L_1, Test.

	Before matching	After matching
Multivariate imbalance measure L_1	.415	.093

Third, a *summary of unbalanced covariates* was generated with univariate diagnostic information for each covariate before and after PSM was used. Before- and after-matching imbalances for covariates can be examined with Cohen's *d*, the Standard Mean Difference values in the results table. Again, before-matching values of Cohen's *d* should not exceed 1, and after-matching $|d| < .25$ is recommended by Stuart and Rubin (2007), or a more conservative benchmark of $|d| < .1$ has become standard (Love, 2008, Shadish et al., 2008). For example, if the after-matching value of Cohen's *d* for a covariate is larger than .25, imbalance occurs in that covariate after the PSM. In current study, no imbalance was found in any of the five covariates.

SPSS also provides standard mean differences for all quadratic and interaction terms of the covariates, if no imbalances exist in the linear terms, the researcher needn't look any further. If, however, there are imbalances that exceed the $|d| > .25$ for post-matching and $|d| > 1$ pre-matching, researchers can use the quadratic and interactive values to determine how they want to manipulate the original data to produce a model without these imbalances, which is the reason SPSS provides this information in the table.

Next, dot plot (see Figure 2) demonstrates the effects of matching on standardized differences of the covariates. In the dot plot, the empty dots are the Cohen's *d*'s for each covariate before matching, and the bold dots are the Cohen's *d*'s for after matching. Notice how the matched covariates have moved closer to 0, which implies matched subjects are more similar than those unmatched subjects on each of the covariates.

Lastly, the common support histograms were used to determine if common support exists between the treated and control subjects' propensity scores after matching, and hence, to ensure that the criteria of common support between matched treatment and control groups is met. In this case (see Figure 3), notice that the matched treated and control subjects have very similar distributions and a great deal of overlap, which is highly desirable for post-matching analyses (Caliendo & Kopeinig, 2008).

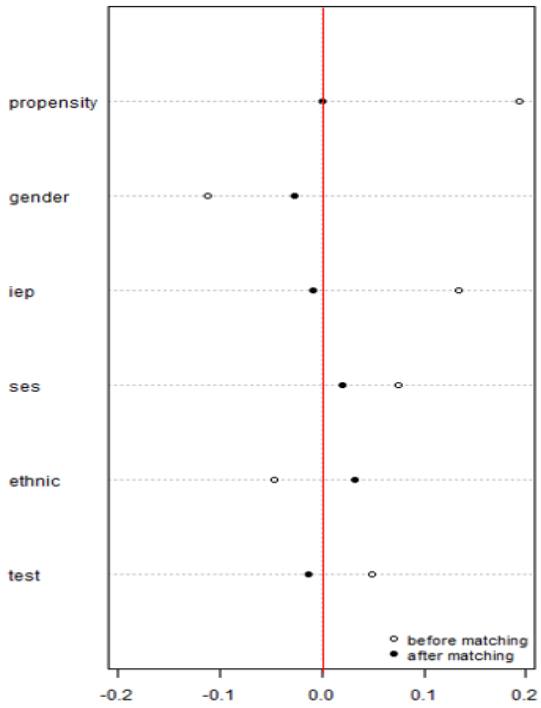

Figure 2: Dot Plot for 1:1 Match Displaying *d* before and after Matching

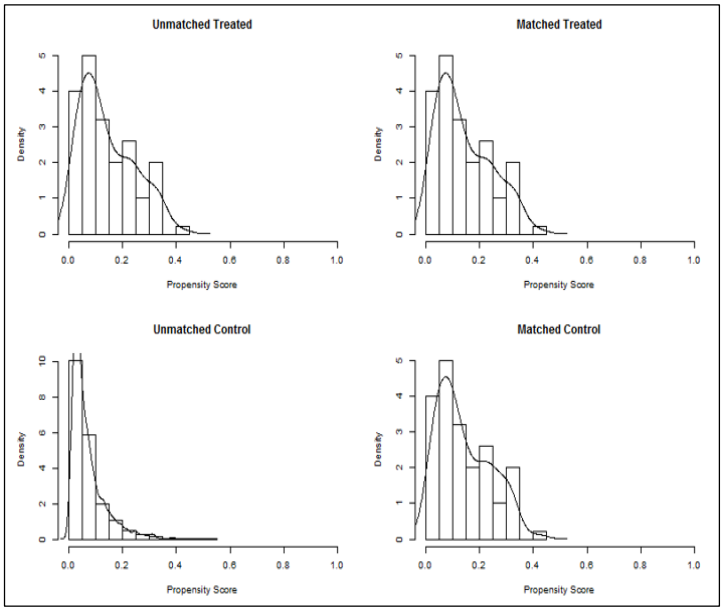

Figure 3: Common Support Histograms for1:1 Matching Method

In summary, based on all tests for imbalance and the graphic illustrated data, the choices of matching methods were confirmed for the PSM in this study. The PSM used Nearest Neighbor matching algorithm, 1:1 treatment-control ratio, and without using replacement and caliper.

On the data set of the 3900 scores, including 300 from IEP treatment group and 3600 from non-IEP control group, the above PSM methods and procedures were performed and

resulted in the 1:1 matched treatment and control groups with an N = 300 for each. The two matched groups were then used for the comparison of mean differences.

DATA ANALYSIS AND RESULTS: t-TEST TO COMPARE MEAN DIFFERENCES

Data Analysis and Results. For the 1:1 matched groups, paired *t*-test was suggested to be one of the preferred methods to estimate mean difference (Austin, 2011; & Imbens, 2004). Three paired *t*-tests were conducted to examine the mean differences on each of the three subscales (*Factual Item Scale*, *Conceptual Item Scale*, or *Scenario Item Scale*) between participants in the IEP program and those who were not in the IEP program (N = 300 pairs). Table 3 shows the descriptive results of the three test scales.

Table 3. Descriptive Results

	IEP Group (n=300)		Non-IEP Group (n=300)	
IV	*M*	*SD*	*M*	*SD*
Factual	5.238	2.035	5.583	2.062
Conceptual	4.855	1.924	5.140	1.988
Scenario	2.733	2.439	3.000	1.409

The results from the paired *t*-tests indicated that students who were not in the IEP program performed better than those who were in the IEP program in their *Factual Item Scale* test scores ($t_{(299)}$ = 2.63, p = .009, d = .15), and *Conceptual Item Scale* test scores ($t_{(299)}$ = 2.38, p = .018, d = .14). There was no difference between the two groups in their *Scenario Item Scale* scores ($t_{(299)}$ = 1.68, p = .094, d = .09).

Rosenbaum and Rubin (1983, 1984) proved that properly matching treatment and control units based on propensity scores, with all assumptions met, will remove more than 90% of any bias present in a study. It is of important reference to analyze the differences in the results between the independent samples *t*-test and the matched-pairs *t*-test. Independent samples *t*-test on the entire set of data before matching was performed, comparing the IEP group and non-IEP group on each of the three measure scales. The results are presented in Table 4. Adjusted α levels with Holm's Sequential Bonferroni Corrections were used for the significance decisions on both sets of tests.

Table 4. Results of Before and After Matching t-Tests.

	After Matching			Before Matching		
IV	*Paired t-test*	*p*	α	*Independent t-test*	*p*	α
Factual	$t_{(299)}$ = 2.63	.009*	.017	$t_{(8738)}$ = 2.16	.030	.025
Conceptual	$t_{(299)}$ = 2.38	.018*	.025	$t_{(8574)}$ = 4.03	.001*	.017
Scenario	$t_{(299)}$ = 1.68	.094	.050	$t_{(8738)}$ = 0.86	.389	.050

Note: α – Adjusted α level with Holm's Sequential Bonferroni Corrections

(*) – significant at adjusted α Level

Notice that both *t*-tests on *Conceptual Items Scale* produced significant differences. However, why was the difference so much greater in the independent samples test (before matching, p<.001) than in the matched-pairs test (after matching, p<.018)? Matching produced a treatment and control group that were balanced based on covariates, and therefore any control subjects that could potentially create greater mean differences between the two groups, without being matched to a treatment subject were removed. This speaks to the inherent bias that is present in observational studies. Theoretically, this produced a more valid outcome to determine differences and treatment effect between the groups. However, when *Factual Items* and *Scenario Items* are examined for differences,

both before and after matching, the *p*-value *decreased* as a result of matching. In the case of the *Factual Items Scale*, the reduction of *p*-value resulted in significant differences after matching ($p<.01$) when there were none before matching ($p<.030$). Again, PSM reduced extant bias present in the data prior to matching, and ensured that experimental subjects were compared to similar control subjects based on balanced covariates. Love (2008) and others recommend sensitivity analyses such as replicating the study with different randomly chosen groups to determine just how much bias is present, and thus removed by matching, between the two groups prior to matching.

Cautions When Interpreting the Results. In this example, two groups were identified and matched with PSM: IEP group and non-IEP group. IEP group represents an existing population consisting of a special group of students who need special individual education services, and non-IEP group represents the general population of students who are in regular education programs. In the above PSM procedures, by simply following the consistent terms used in the PSM literature, we referred to the IEP group as *treatment*, and non-IEP as *control* to identify the groupship. Accurately, they present two different groups, but not treatment-control by nature as in the sense of a pure experimental design. However, the status of IEP can be considered the groupship factor as in an observational context, where PSM applies.

In the PSM procedures described in this study, a set of covariates were used to create the logit model, the covariate balance was carefully examined and all the criteria were met, and then the two propensity-score-matched groups were created. According to Rosenbaum and Rubin (1983, 1984), serious bias (or more than 90% of any bias) present in the study should be removed. Therefore, we may state that the significant results can be considered as the differences by the groupship factor.

SUMMARY AND FURTHER EFFORTS

PSM is a relatively new and innovative statistical method to examine treatment effect for researchers who are using nonexperimental or observational data. Over decades, educational research has relied heavily on quasi-experimental design, for which PSM obviously is a viable statistical tool, as it could provide more options of data analysis, result in relatively more valid research outcomes, and make it possible to use educational data from broader and more diverse resources. This article has introduced the very basics and initial procedures to conduct PSM, and demonstrated the procedures with the data set from a study on technology-integrated science learning. It is the authors' hope that this work could be of reference to educators who are interested in learning and using PSM in their studies.

Further efforts will be made to continually explore PSM methods and applications such as (a) different models of sample selection, (b) PSM matching estimators, (c) propensity score analysis with nonparametric regression, or (d) sensitivity analysis (Guo & Fraser, 2010). The authors are currently developing training materials for the benefit of doctoral students' dissertation studies. Another project done by the authors for further publication is focusing on PSM model selection with test results from a series of covariate balance examinations.

REFERENCES

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate behavioral research, 46*(3), 399-424.

Barth, R. P., Greeson, J. K. P., Guo, S., Green, R. L., Hurley, S., & Sisson, J. (2007). Outcomes for youth receiving intensive in-home therapy or residential care: A comparison using propensity scores. *American Journal of Orthopsychiatry, 77*(4), 497-505.

Blackford, J. U. (2009). Propensity scores: Method for matching on multiple variables in down syndrome research. *Intellectual and Development disabilities, 47*(5), 348-357.

Bertsekas, D., P. (1991). Linear Network Optimizations: Algorithms and Codes. Cambridge, MA. MIT Press.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006), Variable Selection for Propensity Score Models. *American Journal of Epidemiology*, 163, 1149–1156.

Bryson, A. (2002). The union membership wage premium: An analysis using propensity score matching. Discussion Paper No. 530, *Centre for Economic Performance*, London.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31-72.

Chalmers, T. C., Smith, H. Jr, Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., & Ambroz, A. (1981). A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials, 2*(1): 31–49

DaDeppo, L. M. W. (2009). Integration factors related to academic success and intent to persist of college students with learning disabilities. *Learning Disabilities Research and Practice, 24*(3), 122-131.

Dearing, E., McCartney, K., & Taylor, B. A. (2009). Does higher quality early child care promote low-income children's math and reading achievement in middle childhood? *Child Development, 80*(5), 1329-1349.

EOP (Executive Office of the President). (2012). Big data across the federal government. White House. Retrieved November 13 2014 from: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf

George, D., & Malley, P. (2000). *SPSS for Windows Step by Step: A Simple Guide*. Prentice Hall PTR.

Green, S. B., & Salkind, N. J. (2005). *Using SPSS: Analyzing and Understanding Data* (4th ed.). Upper Saddle River, NJ: Pearson, Prentice Hall.

Gujarati, D. N., & Porter, D. C. (2009). *Terminology and Notation. Basic Econometrics* (Fifth international ed.). New York: McGraw-Hill.

Guo, S., & Fraser, M. W. (2010). *Propensity Score Analysis: Statistical Methods and Applications*. Los Angeles: CA: Sage.

Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science, 23*(2), 219-236.

Hitt, L. & Frei, F. (2002). Do better customers utilize electronic distribution channels? The case of PC banking. *Management Science, 48*(6), 732-748.

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

Iacus, S. M., King, G., & Porro, G. (2009). CEM: Coarsened exact matching software. *Journal of Statistical Software*, 30, 1-27.

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

Judge, S., & Watson, S. M. R. (2011) Longitudinal outcomes for mathematics achievement for students with learning disabilities. *The Journal of Educational Research, 104*(3), 147-157.

Krzanowski, W. J. (2000). *Principles of Multivariate Analysis: A User's Perspective*. New York, NY: Oxford University Press.

Lewis, D. (1973). *Counterfactuals*. Oxford. Blackwell Publishers.

Love, T. (2008, January). *Reducing the impact of selection bias using propensity scores*. Presentation delivered at 7th international conference on health policy statistics, Center for Health Care Research and Policy, Case Western Reserve University at MetroHealth Medical Center. Cleveland, Ohio, Retrieved from http://www.chrp. org/propensity/ICHPS2008propensity_love.pdf

McGraw, R., Lubienski, S. T., & Strutchens, M. E. (2006). A closer look at gender in NAEP mathematics achievement and affect data: Intersections with achievement, race/ethnicity, and Socioeconomic status. *Journal for Research in Mathematics Education, 37*(2), 129-150.

Mertler, C. A., & Vannatta, R. A. (2002). *Advanced and Multivariate Statistical Methods*. Los Angeles: Pyrczak

Nieuwbeerta, P., Nagin, D. S., & Blokland, A. A. J. (2009). Assessing the impact of first-time imprisonment on offenders' subsequent criminal career development: A matched samples comparison. *Journal of Quant Criminal, 25*, 227-257.

Porta, M., ed. (2008). *A Dictionary of Epidemiology* (5th ed.). New York: Oxford University Press

Rosenbaum, P. R., Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistics, 39*(1), 33-38.

Rosenbaum, P. R., Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistics Association, 79*(387), 516-524.

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

Rosenbaum, P. R. (2009). *Design Observational Studies*. New York: Springer-Verlag.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*(8), 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.

Segal, J. B., Griswold, M., Achy-Brou, A., Herbert, R., Bass, E. B., Dy, S. M., Millman, A. E., Wu, A. W., & Frangakis, C. E. (2007). Using propensity scores subclassification to estimate effects of longitudinal treatments: An example using a new diabetes medication. *Medical Care, 45*(10), S149-S157.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association*, 103, 1334-1343.

Staff, J., Patrick, M. E., Loken, E., & Maggs, J. L. (2008). Teenage alcohol use and educational attainment. *Journal of Studies on Alcohol and Drugs*, *69*(6), 848.

Stuart, E. A., & Rubin, D. B. (2007). Best practices in quasi-experimental designs: Matching methods for causal inference. *Best Practices in Quantitative Methods*, Chapter 11, Osborne J (ed.) (pp 155-176). Sage Publications: Thousand Oaks.

Steiner, P. M., & Cook, T. D. (2013). Matching and propensity scores. In *The Oxford Handbook of Quantitative Methods*. Oxford: Oxford University Press.

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics, 36*(2), 213-236.

Thoemmes, F. (2012). Propensity score matching in SPSS. arXiv preprint arXiv: 1201.6385.

Wyse, A. E., Keesler, V., & Schneider, B. (2008). Assessing the effects of small school size on mathematics achievement: A propensity score-matching approach. *The Teachers College Record*, *110*(9), 1879-1900.