Using Propensity Scores in Quasi-Experimental Designs to Equate Groups

Forrest C. Lane

Robin K. Henson

University of North Texas

Department of Educational Psychology

Paper presented at the 33rd annual meeting of the Southwest Educational Research Association, New Orleans, LA, February 18, 2010. Correspondence concerning this paper should be addressed to the first author at Forrest.Lane@usm.edu

Abstract

Education research rarely lends itself to large scale experimental research and true randomization, leaving the researcher to quasi-experimental designs. The problem with quasi-experimental research is that underlying factors may impact group selection and lead to potentially biased results. One way to minimize the impact of non-randomization is through the use of propensity scores. First developed by Rosenbaum & Rubin (1983b), these scores allow researchers to balance non-equivalent groups though matching on a singular scalar variable. The present paper will present the theoretical framework behind propensity scores along with a heuristic data set to demonstrate propensity score calculation and evaluation.

Using Propensity Scores in Quasi-Experimental Designs to Equate Groups

Experimental design is historically the only approach to estimating true treatment effects and making causal inferences. This is particularly important in the field of educational research given the growing expectation of Department of Education for increased rigor in program evaluation (Rudd & Johnson, 2008). The problem is that educational research rarely lends itself to large scale experimental design and true randomization (Grunwald & Mayhew, 2008). There are often too many ethical and cost limitations resulting in an overabundance and over-reliance on non-randomized studies throughout the field of education.

This is not to say that findings from non-randomized designs are always inaccurate. In fact, these designs may better reflect the complexity of our educational environment when done properly (Luellen, Shadish, & 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). The causality inferred through experimental design can become too narrowly limited to the context of the conditions within that experiment.

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. Non-randomized groups may systematically differ from one another based on any number of covariates (Rosenbaum & Rubin, 1983a). For instance, a researcher might be interested in impact of advanced high school curriculum on college success or the effectiveness of co-curricular programs and services on student development. However, students may participate in these programs based on any number of theoretically relevant variables. This can lead to bias

when interpreting treatment effects when pre-group differences have not been accounted for in the research design (Grunwald & Mayhew, 2008).

As an applied example, Shadish, Luellen, and Clark (2006) conducted a study to empirically demonstrate the problems due to non-randomization. In their study, participants were randomly assigned to experimental and quasi-experimental conditions in order to assess the gains from math and verbal training on quantitative and communication skills. Coincidentally, quasi-experiment participants performed better in all treatment conditions than those in the randomized experiment. Students who chose to be in the math condition group and received a math treatment performed better than those who would have randomly assigned to the same group in the randomized experiment, thus demonstrating the potential bias of quasi-experimental research designs.

Several methods have been employed over the years to accommodate problems of endogeneity. For example, researchers have used regression analyses or structure equation modeling (Titus, 2006). However, these designs can never fully control for all potential background variables nor is there a consensus for how to compute adjustment coefficients (Grunwald & Mayhew, 2008; Shadish, Luellen, & Clark, 2006). As such, the potential for selection bias is increased and researchers are subject to treatment effects which may be confounded by group differences as a result of non-randomization. This limits one's ability to accurately report treatment effects and make causal inferences (Hong & Raudenbush, 2005).

In response to the challenges of quasi-experimental designs, fields such as medicine, statistics, and economics have been using propensity score matching as a method to control for group differences when estimating treatment effects (D'Agostino, 1998; Grunwald & Mayhew, 2008; Shadish, Luellen, & Clark, 2006). This method uses a participant's probability of group

membership as a scalar variable to balance participants. The advantage of this technique is that it takes into account the complexity of educational environments while providing the kind of rigor needed in an evidence-based assessment climate.

Educational researchers should be more informed of this analysis because propensity score matching is a recommended method by the U.S. Department of Education to improve the strength of quasi-experimental research. However, amidst the calls for more scientifically based methodology within education, propensity score matching remains greatly underutilized in the literature (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007; Slavin 2009). The purpose of this paper is to demonstrate propensity score calculation and matching using a heuristic example. Specifically, the theoretical framework behind the analysis will be presented along with a demonstration of the calculation of propensity scores through the use of logistic regression and nearest neighbor matching within calipers.

Theoretical Framework

The literature on use of propensity scores can be attributed to the seminal work of Rosenbaum & Rubin, (1983b). Their aim was to address limitations of non-randomization by deriving a mathematical solution to account for group differences. In true randomization, we would expect an equal likelihood or probability of group membership. For example, if we were to flip a coin we should expect to receive heads approximately half the time and the other half tails. Our probability of getting either outcome is 50% and would result in a propensity score of $p = .50$. Similarly in experimental designs, participants are randomly assigned and have an equal likelihood of being in either a control or treatment condition. Groups can be directly compared to one another because systematic differences have been controlled through experimental design.

Conversely, quasi-experimental designs are subject to participant self-selection which introduces bias when comparing groups. Using the same coin analogy, there is an unequal probability of group assignment and this likelihood is not known ($p \neq .50$). This probability must be estimated which is accomplished using covariates in the analysis to calculate a probability value.

A propensity score (e) for an individual (i) is defined as the conditional probability (P) of assignment to a particular treatment or control group (T) given a set of covariates (X) (Rosenbaum & Rubin, 1983), expressed as,

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

Theoretically relevant pretreatment variables are used to derive probabilities of group membership and then used to match participants across groups (Rosenbaum & Rubin, 1983b). Once matched, treatments effects should be more reflective of the true effect and analogous to interpretation of randomized designs.

Traditional matching methods have been laborious, especially as the number of those covariates increases. Propensity score matching minimizes this problem by incorporating covariates into a singular scalar variable ranging from 0 to 1. Once calculated, this new scalar score can be used to balance control and treatment groups though matching, stratification, or regression adjustment which allows for participants in both groups to be equated to one another, simulating the characteristics of randomized studies and serving as a better measure of treatment effects (Austin, 2008; D'Agostino, 1998; Shadish, Luellen, & Clark, 2006).

The benefits of propensity score matching have been demonstrated throughout the literature. For example, Grunwald & Mayhew (2008) conducted a study in the development of moral reasoning in young adults and demonstrated a significant reduction is the overestimation

of effect size from the model. Morgan (2001) used propensity score matching to demonstrate the effect of private school education on math and reading achievement is actually larger than findings in non-matched samples (Schnider et. al, 2007). Other similar studies have been demonstrated in economics (Dehejia & Wahba, 2002), medicine (Schafer & Kang, 2008; Austin, 2007), and sociology (Morgan & Harding, 2006). Although propensity score matching continues to be demonstrated as a superior quasi-experimental method in the literature, it remains underutilized in educational research.

Propensity Score Matching

Covariate Selection

In non-randomized studies, group selection can be influenced by a number of covariates leading to effect size bias. Therefore, theoretically relevant covariates which are 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 which can be used in propensity score matching. Previous literature suggests that any covariate improving predictability ought to be included and not limited to the most parsimonious explanation. However, researchers should seek to identify those covariates which are likely to influence treatment selection, grounded in literature, and thus provide a more meaningful and statistical approximation of group membership (Rubin & Rosenbaum, 1983b; Rubin & Rosenbaum 2002).

Propensity Score Estimation

The probability of group membership or propensity scores calculated for participants in both control and treatment groups. The most commonly used methods include using either logistic or probit regression but can include classification trees or ensemble methods such as bagging, boosted regression trees, and random forest (Austin, 2008; Shadish, Luellen, & Clark,

2006). Classification trees apply classification algorithms help to select variables, identify interactions, and automatically supply strata. This can be advantageous to other methods, such as logistic regression, but may have a tendency to over fit increasing the misclassification rate. Alternatively, ensemble methods allow for the creation of multiple classification trees using subsamples of the initial data set. However, logistic regression is the most widely used method in the literature and will be demonstrated in the following example. It is also the easiest approach for interpreting propensity scores given they are the predicted group probabilities.

Matching & Model Evaluation

Once individuals in both control and treatment groups have been assigned a probability of group membership, participants are matched across groups on their likelihood of being assigned to either condition. The goal is to produce groups in which the same distribution of propensity scores exist for both treated and control units (Rosenbaum & Rubin, 1984). Three primary methods exist for achieve balance including matching, regression adjustment, and stratification, (D'Agostino, 1998). Matching controls for covariates by pairing participants across treatment groups. This is accomplished by 1) matching a participant on the nearest possible propensity score, 2, matching with in calipers, 3) Mahalanobis metric matching, or 4) Mahalanobis metric matching on specified distance based on the average of the variances within treatment groups.

Alternatively, researchers may use an adjustment in the regression analysis. Two methods can be employed in approach including subtracting the effect of covariates from the treatment effect or by adding the propensity score as a variable in the regression equation as an adjustment to the treatment effect. Lastly, stratification (also called sub-classification) can be used which groups participants into equal strata so that participants can be compared based on

the strata they are assigned. Stratification across quintiles has been shown to reduce approximately 90% of bias due to covariates (Rubin & Rosenbaum, 1983b; Rubin & Rosenbaum, 1984; Shadish, Luellen, & Clark, 2005).

Balance in the newly matched sample can be validated through numerical and graphical summaries of the data. Numerical methods on continuous covariates include t-tests while χ^2 tests can be used for nominal or ordinal level covariates (Grunwald & Mayhew, 2008). Rosenbaum & Rubin (1984) recommend testing covariates in a 2 x 5 two-way ANOVA (treatment & control) when stratifying across quintiles to assess the magnitude and significance of group assignment along with interaction effects. Groups are assumed to be balanced when *F*-values are small and there are no significant interaction effects (Rubin 2002).

Once the model is balanced and participants across groups have been matched based on propensity scores, treatments effects can then be estimated on the outcome variable(s) by comparing newly match treatment and control group samples through either a t-test or multi-group equivalent. It's important to note that propensity score matching assumes that, once balanced, there are no systematic differences between groups. Therefore, groups can be directly compared and causal inferences inferred.

Heuristic Example

Hypothetical Scenario

Living learning communities (LLC) are floors or entire residence halls designed to meet the needs of students with common academics, social and cultural interests. Recent literature has suggested that participation in a Living Learning Community (LLC) contributes to improved 1st year academic performance (GPA) and retention (Zhao & Kuh, 2004; Hotchkiss, Moore & Pitts,

2006). Therefore, University X has recently implemented a LLC program at the institution and is interested in assessing the effects of the program. The program is open to all students who choosing to participate. However, facilities are limited and thus a competitive application process was implemented in selecting students for participation in the program.

Sample

A small heuristic sample ($N = 30$) was developed for the purposes of illustrating the analysis. Both control (non-LLC) and treatment (LLC) participants were created to evaluate the program's success based on an inferred quasi-experimental design. Self-selection into the program was assumed which has been shown in the literature to lead effect size bias. Therefore, several theoretically relevant variables were identified in order to better match control and treatment group participants so that true treatment effects could be estimated. Specifically, five predictor variables were identified as based on relevant literature and institutional research (Table 1). Prior academic performance is thought to play a role in LLC participation (Pasque & Murphy, 2005) and so SAT and PSAT scores were simulated and used in the analysis. Additionally, university attachment is thought to contribute to student engagement and program participation. Therefore, two dichotomous variables were created to measure university attachment including residency (in-state vs. out-of-state) and parental affiliation (Alumni vs. Non-Alumni). Lastly, women have been shown to be more engaged on college campuses and so gender was also included as a covariate.

Table 1

Hypothetical Data Set & Covariates of Participation in a Living Learning Community (N=30)

ID	LLC	Instate	Alumni	PSAT	SAT	Gender	GPA	Propensity
1	0	0	1	599	684	0	2.90	0.212
2	0	1	0	602	483	1	3.10	0.460
3	0	0	0	710	422	0	2.85	0.179
4	0	0	0	452	586	0	2.75	0.048
5	0	1	1	578	628	1	3.25	0.642
6	0	1	0	687	768	0	3.45	0.472
7	0	1	0	679	785	1	3.50	0.510
8	0	0	0	642	675	0	3.35	0.114
9	0	0	1	668	677	0	3.60	0.279
10	0	1	1	685	696	1	3.60	0.750
11	0	0	1	765	747	0	3.75	0.381
12	0	1	1	490	736	1	3.21	0.512
13	0	0	0	530	430	0	2.75	0.078
14	0	1	1	622	423	0	2.95	0.679
15	0	1	0	762	748	0	3.45	0.574
16	0	0	1	575	575	0	3.05	0.203
17	1	1	0	758	772	1	3.85	0.614
18	1	1	1	641	749	0	3.75	0.652
19	1	0	0	662	567	0	3.25	0.133
20	1	1	1	510	702	1	3.33	0.544
21	1	1	1	474	614	0	3.05	0.461
22	1	1	1	560	646	0	3.25	0.568
23	1	1	1	592	720	1	3.35	0.645
24	1	1	0	773	703	1	3.85	0.643
25	1	1	1	600	800	0	4.00	0.594
26	1	1	0	747	441	1	2.90	0.653
27	1	1	1	704	795	1	3.45	0.756
28	1	0	1	732	534	0	3.10	0.374
29	1	1	1	652	460	1	3.25	0.748
30	1	1	0	680	720	1	3.75	0.522
M				642.87	0.43		3.32	0.467
SD				122.70	0.50		0.34	.216

Statistical Software

No commercially available programs exist to conduct propensity score matching. However, propensity score calculation can be conducted in SAS, STATA, SPSS, and R and

syntax is generally available in the literature to perform the matching analysis. For example, the PSMATCH2 algorithm is available in STATA (Leuven & Sianesi, 2004), the SUGI 214-26 "GREEDY" Macro in SAS (D'Agostino, 1998), and an SPSS algorithm on the University of North Carolina's Jordan Institute for Families website (Painter, 2009). SPSS v18.0 and syntax will be illustrated in this example since it tends to be program with the greatest familiarity among social science researchers. Specifically, Painter's (2009) syntax was used (Appendix A) and modified for matching with in calipers in order to fit the data below.

Propensity Score Estimation & Matching

The effect of academic performance by group was evaluated prior to matching using an independent samples t-test. Results indicated that that LLC participants generally outperformed non-participants in terms of GPA at the end of one year (Table 2). The results were not statistically significant ($t(28) = 1.795$, $p = .084$), likely a function of the small sample used in the analysis. However, the quarter point difference in GPA was determined to be practically significant given the standardized mean difference between the two groups ($d = .660$) was large using guidelines provided by Cohen (Hinkle, Wiersma, Jurs, 2003). This suggested that participation in an LLC was practically significant in explaining differences in academic performance.

Table 2

Independent T-Tests Results Prior to Propensity Score

	N	*M*	*SD*	*t*	*df*	*p*	*d*
Academic Achievement							
Non-LLC Participants	16	3.21	.323	1.795	28	.084	.660
LLC Participants	14	3.43	.343				
Propensity Scores (Bias)							
Non-LLC Participants	16	.380	.226	2.534	28	.017	.942
LLC Participants	14	.565	.161				

Propensity scores were then calculated using logistic regression with the five covariates discussed above to assess pre-group differences due to self-selection. Predicted probabilities (propensity scores) were saved as a part of the analysis (Table 1). An independent samples *t*-test was conducted to examine the initial bias using the propensity scores (table 2). Results indicated a substantial bias in the likelihood between the groups towards participation in an LLC ($t(28) = 2.534$, $p = .017$, $d = .942$). In fact, the difference was nearly an entire standard deviation and larger than effect found between the two groups on the outcome variable (GPA) prior to matching. This suggests that the two groups should not be directly compared when estimating treatment effects.

In order to equate the two groups, SPSS syntax (Appendix A) was used to match participants on the nearest propensity score. All treatment cases were matched, leaving two unused cases from the control group ($N = 28$). The distance was then calculated between each matched pair of control and treatment cases using the logit transformation of the propensity score, or the standardized predicted value obtained from the logistic regression. Matched pairs were either kept for further analysis or discarded using calipers. A caliper of 0.25 standard deviations has been shown to substantially reduce bias and can be effective as a matching technique (Rosenbaum and Rubin, 1985b; Stuart & Rubin, 2007). Therefore, only matched pairs with a distance of less than 0.25 were retained for further analysis.

The first seven matched pairs met the criteria set forth by Stuart & Rubin (2007). Participants 14 and 17 had a standardized mean distance of .26 but considered close enough due to rounding differences and thus retained. The remaining matched samples had standardized differences ranging between .46 and 2.41 standard deviation and were not kept for further analysis. This resulted in eight (8) matched pairs of control and treatment participants ($N = 16$).

Table 3

*Nearest Neighbor Matching within Calipers**

Control ID	Propensity Score	Logit Score	Treatment ID	Propensity Score	Logit Score	*d* (Caliper)
8	.133	-1.87	19	.114	-2.05	.16
11	.374	-.52	28	.381	-.49	-.03
2	.461	-.16	21	.460	-.16	.00
12	.522	.09	30	.512	.05	.04
15	.544	.18	20	.574	.30	-.11
7	.568	.27	22	.510	.04	.21
5	.594	.38	25	.642	.58	-.19
14	.614	.46	17	.679	.75	-.26
10	.643	.59	24	.750	1.10	-.46
6	.645	.60	23	.472	-.11	.64
9	.652	.63	18	.279	-.95	1.43
1	.653	.63	26	.212	-1.32	1.77
16	.748	1.09	29	.203	-1.37	2.23
3	.756	1.13	27	.179	-1.52	2.41
4	.048	-2.99	-	-	-	-
13	.078	-2.47	-	-	-	-

*Standardized distance measures were obtained using logit scores of participants and a pooled *SD* = 1.10067.

The newly matched sample was then subjected to an independent samples *t*-test to determine if the matching had reduced the bias to a sufficient level to allow for further comparison. The standardized difference in the mean propensity score between the two groups should be near zero ($d < .20$) in order to be considered balanced (Rubin, 2001). Results showed a 95% bias reduction due to covariates based on matching within calipers. Group differences had now been reduced to both statistically and practically non-significant results ($t(14) = .092$, $p = .928$, $d = .047$) and were within guidelines established in the literature. Should bias not have been sufficiently reduced, additional covariates could have been included in the generation of propensity scores until an adequately matched sample could be obtained.

Once balance has been achieved, groups can then be compared directly using a *t*-test or multi-group equivalent on the outcome of interest. Any differences found as a result of matching should be reflective of the true treatment effect and analogous to experimental design. Thus, an independent samples *t*-test was then performed on GPA to re-examine the effects of participation in an LLC. Results indicated that the effects of participation were much smaller (41%) than initial estimates ($t(14) = .816$, $p = .428$, $d = .384$). Although LLC participation may still produce practical differences in academic performance, results from the matched sample suggested that these effects were not as strong as initially identified prior to matching.

Table 4

Independent T-Tests Results Post Propensity Score

	N	*M*	*SD*	*t*	*df*	*p*	*d*
Academic Achievement							
Non-LLC Participants	8	3.32	.249	.816	14	.428	.384
LLC Participants	8	.3.44	.364				
Propensity Scores (Bias)							
Non-LLC Participants	8	.484	.171	.092	14	.928	.047
LLC Participants	8	.476	.158				

Discussion

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 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, Shadish, & Clark, 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. This was reduced by as much a 95% in the present example, illustrating both the robustness of this matching 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 in to meet the demands of a growing assessment-based climate.

Several considerations should also be discussed prior to utilizing to propensity score matching. First, propensity score matching has been questioned regarding it benefits relative to more familiar methods such as ANCOVA. There is a fundamental difference between controlling for variables which may contribute to differences on an outcome variable versus those as a result of self-selection and non-randomization (Miller & Chapman, 2001). Only after participants in both treatment groups have been matched on their propensity score should ANCOVA be considered as an appropriate technique to control group differences on an outcome variable of interest. Additionally, many matching programs are readily available within the literature but each make varying assumptions regarding the criteria for participant matching. No consensus was found within the literature with regard to how these various algorithms impact results. It's assumed that these methods produce similar bias reduction but readers should be cautioned that there is little evidence to support this conclusion. Finally, propensity score matching will likely yield smaller samples than initially obtained in the data collection process. Matching can be conducted with replacement, increasing retained control participants. However, smaller samples are almost inevitable. Therefore, sampling adequacy for statistical power should be considered *a priori* to conducting the analysis.

References

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.

Glenn, D. (2005, March). New federal policy favors randomized trials in education research. *The Chronicle of Higher Education*

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika, 95*, 481-488. dio:10.1093/biomet/asn004.

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

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.

Hotchkiss, J. L., Moore, R. E., & Pitts, M. M. (2006). Freshman learning communities, college performance, and retention. *Education Economics, 14,* 197-210.

Freedman, D. A., (2008). Weighting regressions by propensity scores. *Evaluation Review, 32*, 392-409. doi: 10.1177/01938X08317586.

Leow, C., Marcus, S., Zanutto, E., & Boruch, R. (2004). Effects of advanced course-taking on math and science achievement: Addressing selection bias using propensity scores. *American Journal of Evaluation, 25,* 461-478.

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.

Miller, G. A. & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110,* 40-48.

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

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

National Research Council (2000). *Scientific research in education*. Washington, D.C.: National Academy Press.

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

Pasque, P. A. & Murphy, R. (2005). The intersections of living-learning programs and social identity as factors of academic achievement and intellectual engagement. *Journal of College Student Development*, 46, 429-441.

Pike, G. (2009). The differential effects of on- and off-campus living arrangements on students' openness to diversity. *Journal of Student Affairs Research & Practice, 46*, 629-645.

R Development Core Team. (2006). *R: A language and environment for statistical computing*. Vienna, Austria.

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

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

Rosenbaum, P. R. (1996). Observational studies and nonrandomized experiments. In S. Ghosh & C. R. Rao (Eds.), *Handbook of statistics* (Vol. 13, pp. 181-197). Amsterdam: Elsevier Science B.V.

Rosenbaum, P. R. (2001). Using propensity scores to help design observational studies: Application to tobacco litigation. *Health Services & Outcomes Research Methodology* 2, 169-188.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. Health Services & Outcomes Research Methodology 2, 169–188.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52,* 249-264.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. Journal of the American Statistical Association, 95, 573-585.

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,* 180-188.

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.

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: Bootzin R.R., McKnight P.E. (Eds.), *Strengthening research methodology: Psychological measurement and evaluation*. American Psychological Association: Washington, DC, 143–157.

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

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.

Titus, M.A. (2006). No college student left behind: The influence of financial aspects of a state's higher education policy on college completion. *The Review of Higher Education, 29,* 293-317

Whitehurst, G. (2002). http://www.ed.gov/nclb/methods/whatworks/eb/edlite-slide021.html.

Zhao, C. & Kuh, G. (2004). Adding value: Learning communities and student engagement. *Research in Higher Education, 45*, 115-138.

PASW (v17.0) Syntax for Propensity Score Matching using Matching within Calipers

```
DATASET ACTIVATE DataSet1.
**************************************************************************
*Initial logistic regression to compute propensity scores, indicated in the dataset as 'propen').
**************************************************************************

LOGISTIC REGRESSION VARIABLES LLC
  /METHOD=ENTER Instate Alumni PSAT SAT Gender
  /SAVE=PRED
  /CLASSPLOT
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
RENAME VARIABLES (PRE_1=propen).

**************************************************************************
*The logit of the propensity scores are then calculated.  "Rosenbaum and Rubin (1985b) suggest
that if the caliper matching is done using the propensity score, the bias reduction is obtained on
all of the covariates that went into the propensity score. They suggest a caliper of 0.25 standard
deviations of the logit transformation of the propensity score can work well in general" (Stuart &
Rubin, 2007, ¶4.3.3).
**************************************************************************

Compute lnprop=LN(propen/(1-propen)).
execute.

**************************************************************************
*Calculation of the mean and standard deviation of logit scores.  This will be used to estimate a
.25 caliper later in the syntax.
**************************************************************************

DESCRIPTIVES VARIABLES=lnprop
  /STATISTICS= MEAN STDDEV.

**************************************************************************
*Initial T-test results comparing GPA groups to demonstrate how results may be mistakenly
interpreted in quasi-experimental designs when propensity scores are not taken into account.
**************************************************************************
T-TEST GROUPS=LLC(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=Achievement
  /CRITERIA=CI(.95).
**************************************************************************
```

```
*T-test between treatment groups based on propensity score to demonstrate the initial bias due to
covariates.
*****************************************************************************

T-TEST GROUPS=LLC(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=propen
  /CRITERIA=CI(.95).

*****************************************************************************
*The following syntax indicates where the file can be obtained in your directory.  Change the
path to desired location after ().
*****************************************************************************

DEFINE !pathd() 'F:\Propensity Scores\Forrest Lane Example' !ENDDEFINE.

FREQUENCIES
  VARIABLES=LLC
  /ORDER=  ANALYSIS.

SAVE OUTFILE=!pathd + "population.sav" .

*****************************************************************************
* Core code written by Raynald Levesque */
* Adapted for use with propensity matching by John Painter Feb 2004*/
* Program developed and tested with SPSS 11.5 */
* Procedure will find best match for each treatment case from the control cases, */
* control case is then removed and not reconsidered for subsequent matches */
* Order of cases is randomized */OUTFILE.

* Requirement: The number of Treatment cases must be known */

* Change file path here only  */OUTFILE.


********************* .
** End Preparation .
********************* .
GET FILE= !pathd + "population.sav".
COMPUTE x = RV.UNIFORM(1,1000000) .
SORT CASES BY LLC(D) propen x.
COMPUTE idx=$CASENUM.
SAVE OUTFILE=!pathd + "mydata.sav".

* Erase the previous temporary result file, if any.
ERASE FILE=!pathd + "results.sav".
```

```
COMPUTE key=1.
SELECT IF (1=0).
* Create an empty data file to receive results.
SAVE OUTFILE=!pathd + "results.sav".
exec.

*********************************************************************************
* Syntax below defines the a macro which will do the job.
*********************************************************************************

SET MPRINT=no.
*//////////////////////////////.
DEFINE !match (nbtreat=!TOKENS(1))
!DO !cnt=1 !TO !nbtreat

GET FILE=!pathd + "mydata.sav".
SELECT IF idx=!cnt OR LLC=0.
* Select one treatment case and all control .
DO IF $CASENUM=1.
COMPUTE #target=propen.
ELSE.
COMPUTE delta=propen-#target.
END IF.
EXECUTE.
SELECT IF ~MISSING(delta).
IF (delta<0) delta=-delta.

SORT CASES BY delta.
SELECT IF $CASENUM=1.
COMPUTE key=!cnt .
SAVE OUTFILE=!pathd + "used.sav".
ADD FILES FILE=*
          /FILE=!pathd + "results.sav".
SAVE OUTFILE=!pathd + "results.sav".

*********************************************************************************
*Match back to original and drop case from original .
*********************************************************************************

GET FILE= !pathd + "mydata.sav".
SORT CASES BY idx .
MATCH FILES
 /FILE=*
 /IN=mydata
 /FILE=!pathd + "used.sav"
 /IN=used
```

```
 /BY idx .
SELECT IF (used = 0).
SAVE OUTFILE=!pathd + "mydata.sav"
 / DROP = used mydata key delta.
EXECUTE.
!DOEND
!ENDDEFINE.
*//////////////////////////////.

SET MPRINT=yes.

*************************************************************************.
* MACRO CALL (first insert the number of treatment group cases after nbtreat below) .
*************************************************************************.
!match nbtreat=14.

* Sort results file to allow matching.

GET FILE=!pathd + "results.sav".
SORT CASES BY key.
SAVE OUTFILE=!pathd + "results.sav".

*************************************************************************.
* Match each treatment cases with the most similar non treatment case.
* To include additional variables from original file list them on the RENAME subcommand
below .
*************************************************************************.

GET FILE=!pathd + "mydata.sav".
MATCH FILES
 /FILE=*
 /FILE=!pathd + "results.sav"
 /RENAME (idx = d0) (id=id2) (propen=propen2)  (LLC=LLC2) (Instate=Instate2)
(Alumni=Alumni2) (PSAT=PSAT2) (SAT=SAT2) (Gender=Gender2) (lnprop=lnprop2)
(Achievement=Achievment2) (key=idx)
 /BY idx
 /DROP= d0 x.
FORMATS delta propen propen2 (F10.8).
SAVE OUTFILE=!pathd + "mydata and results.sav".
EXECUTE .

**************************************************************************
*This is an evaluation of the difference between matched pairs.  Matching was done on nearest
neighbor matching.
```

```
*However, a caliper of .25 standard deviations of the logit of the propensity scores was used to limit matches used to only those matches which had a distance of <.25 standard deviations of logit scores.
**********************************************************************************

COMPUTE CALIPER25=((lnprop-lnprop2)/1.23).
EXECUTE.

**********************************************************************************
*The new data set of the eight matched pairs (participants with caliper <.25) was saved into a new data set.
* An independent sample T-Test was conducted on the newly matched sample to examine for statistical differences between the groups based on propensity score.
**********************************************************************************
T-TEST GROUPS=LLC(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=propen
  /CRITERIA=CI(.95).

**********************************************************************************
 An independent sample T-Test was then conducted on the new matched sample to examine for statistical differences on the outcome variable.
**********************************************************************************
DATASET ACTIVATE DataSet2.
T-TEST GROUPS=LLC(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=Achievement
  /CRITERIA=CI(.95).
```