Abstract Title Page
Not included in page count.

Title:

Using Propensity Score Methods to Approximate Factorial Experimental Designs[*]

Author(s):

Nianbo Dong, Ph.D.
Peabody Research Institute
Vanderbilt University

nianbo.dong@vanderbilt.edu

[*] Updated on July 15, 2011. The author thanks Mark Lipsey and two anonymous reviewers for their valuable comments on previous versions. All errors remain the author's. Any comments/suggestions would be appreciated.

Abstract Body

Background / Context:

When randomized experiments are not feasible, propensity score methods can be applied to approximate randomized experiments. The propensity score is the conditional probability for a study unit to receive treatment given a vector of covariates, and propensity score methods can produce unbiased estimate of treatment effect if there are no unmeasured confounders (Rosenbaum, 2002; Rosenbaum & Rubin, 1983). Propensity score methods have been widely applied to approximate experiments with one factor with two levels (e.g., treatment vs. control groups). The propensity scores can be estimated using binary logistic regression. The estimated propensity scores can be further used to estimate treatment effects through many conventional approaches, e.g., subclassification, matching, as covariate, or as weighting function, etc. (see Steiner & Cook (in press) for a review). Recently propensity score methods were generalized to analyze one factor with multiple levels (>2). The generalized propensity scores can be estimated as the conditional probability of receiving a particular level of treatment using polytomous choice model (e.g., multinomial or ordinal logit model) (Imbens, 2000). The resulting estimated propensity scores are then used in subsequent analyses by weighting (inverse of propensity scores) (Imbens, 2000), or optimal nonbipartite matching (Lu, Greevy, Xu, & Beck, 2011; Lu & Rosenbaum, 2004). Furthermore, Imai & van Dyk (2004) generalized propensity score methods to a continuous treatment variable and bivariate treatment variables, and applied the estimated propensity score to subclassify sample to estimate the treatment effects.

Recently researchers and policy makers had increased interested in knowing the effects of multiple factors, particularly the interaction effects of these multiple factors. For example, people not only want to know the main effects of the new math curriculum and high quality teachers, respectively, they also want to know if the new math curriculum taught by the high quality teachers has better effects than taught by the low quality teachers, i.e., the interaction effect. For another example, people want to know the effects of both Head Starts (as compared with other center-based care) and the child care quality on child outcomes, as well as if children in high-quality Head Starts have better outcomes than their counterfactuals (i.e., if they were in high-quality other center-based care).

Ideally, we can use a 2×2 factorial design (Figure 1 in Appendix B), i.e., randomly assign students to four groups, to estimate the main and interaction effect. However, when the full random assignment is not feasible, e.g., students could be randomly assigned to new math curriculum and business as usual groups, but could not be randomly assigned to different teacher quality groups, or no any random assignment is feasible, the four cells (groups) in Figure 1 may be systematically different from each other. The challenge is how we can get unbiased estimates of the main and interaction effects of these two factors.

Researchers have applied propensity score methods to analyze two factors, e.g., analyzing the impact of one program (Factor A) on subgroups (Factor B) (see Hill, Brooks-Gunn, & Waldfogel, 2003; Lochman, Boxmeyer, Powell, Roth, & Windle, 2006; Peck, 2003; Schochet & Burghardt, 2007). The basic practice is to match participants between two levels of Factor A within each of two levels of Factor B, e.g., matching participants between treatment (program) and control (comparison) groups within dose subgroups, or matching participants between high-dose and low-dose groups within treatment and/or control groups. These separate matches may make baseline equivalent between two levels of the matched factor, hence it may produce unbiased estimates of main effects of that factor, however, without making efforts to make

baseline equivalent among all four groups, it may not produce unbiased estimate of the interaction effect.

Imai & van Dyk (2004) have generalized propensity score methods to study bivariate treatment variables (two continuous treatment variables) using stratification, however, less studies on applying propensity score methods to analyze two factors have been conducted so far.

Purpose / Objective / Research Question / Focus of Study:

The purpose of this study is through Monte Carlo simulation to compare several propensity score methods in approximating factorial experimental design and identify best approaches in reducing bias and mean square error of parameter estimates of the main and interaction effects of two factors.

Significance / Novelty of study:

Previous studies focused more on unbiased estimates of the effects of one factor, or the effects of one factor by the subgroups of another factor. The approaches for the unbiased estimates of the main and interaction effects of two factors in studies without full randomization were less examined. The current study will identify appropriate propensity score methods to analyze multiple factors, in particular, the interaction effects.

Research Design:

This paper first reviews Imai & van Dyk's (2004) approach to analyzing two continuous treatment variables through stratifying sample based on the propensity scores on these two treatment variables. We extended Imai & van Dyk's (2004) approach to analyzing two factors (binary variables) using stratification and matching method based on the propensity scores on these two factors. We proposes a new application of propensity scoring, "factorial propensity score matching method", in analyzing two factors. This is followed by introducing other potential propensity score methods in analyzing two factors. Finally, a Monte Carlo Simulation is used to evaluate these propensity score methods.

One of the great contributions in Imai & van Dyk's (2004) was to generalize propensity score application to analyzing two continuous treatment variables. Two propensity score functions for two continuous treatment variables are estimated based on two independent Gaussian linear regression models. The estimated two propensity score functions are used to subclassify the data into several subclasses. Figure 2 illustrates 3×3 subclassification based on two propensity score functions (Imai & van Dyk, 2004). Data are subclassified into three subclasses (lower third, middle third, and upper third) based on each of two propensity score functions, respectively. Each cell of the 3×3 table represents a subclass based on two propensity score functions jointly. In addition to 3×3 subclassification, Imai & van Dyk's (2004) also presented simulation results for 2×2 and 4×4 subclassification. In general, more subclasses produce less bias. Note that the subclassification method based on two propensity score functions could be used for two binary treatment variables. In addition, we could match data based on two propensity score functions.

<u>"Factorial Propensity Score Matching Method"</u>

This is not a new algorithm for propensity score matching, but an application of currently available matching algorithm in analyzing the effects of two factors. The essential idea of this method is to create four equivalent groups to approximate a 2×2 factorial experiment design.

Suppose there is a study with two factors: Factor A with two levels and Factor B with two

levels, like Figure 1. However, there is no full randomization, i.e., participants are not randomly assigned to four groups. The purpose of the study is to examine both the main effects and interaction effects of Factors A and B. The procedures of "factorial propensity score matching method" are below (Figure 3):

1. Estimating two independent propensity score functions based on Factor A and Factor B, respectively, using binary logistic regression model
   1.1 Estimating propensity scores of being in Level 1 of Factor A
   1.2 Estimating propensity scores of being in Level 1 of Factor B
   1.3 Obtaining common support sample by including sample among four groups with overlapping propensity scores for Factor A and Factor B.
2. Matching data using propensity scores based on Factor A within each of two levels of Factor B using the common support sample
   2.1 Within Level 1 of Factor B, matching A1B1 with A2B1 using existing matching algorithm, e.g., optimal matching (Ming & Rosenbaum, 2001), or greedy matching. The matched participants are in yellow and green areas in Cells A1B1 and A2B1.
   2.2 Within Level 2 of Factor B, matching A1B2 with A2B2 using existing matching algorithm. The matched participants are in yellow and green areas in Cells A1B2 and A2B2.
   2.3 Combining two matched datasets (named "matched on A", in yellow and green areas)
3. Matching data using propensity scores based on Factor B within each of two levels of Factor A using the common support sample
   3.1 Within Level 1 of Factor A, matching A1B1 with A1B2 using existing matching algorithm. The matched participants are in blue and green areas in Cells A1B1 and A1B2.
   3.2 Within Level 2 of Factor A, matching A2B1 with A2B2 using existing matching algorithm. The matched participants are in blue and green areas in Cells A2B1 and A2B2.
   3.3 Combining two matched datasets (named "matched on B" (in blue and green areas))
4. Finding the common participants in two matched datasets. This dataset consisting of common participants serves as the final analysis sample (in 4 green areas).

Through these steps we can obtain four matched groups for impact analysis.

Both the "factorial propensity score matching method" and Imai & van Dyk's (2004) subclassification requires estimating two independent propensity score functions based on Factors A and B, respectively. Alternatively, we can collapse two dimensions (2×2) of treatments to one dimension (1×4) of treatment and use the propensity score methods that are appropriate for one-dimensional multiple groups analysis to analyze the effects of two factors.

Other Propensity Score Methods for Analyzing Multiple Groups

First, we convert 2×2 design to 1×4 design, i.e., a design having one treatment variable with four levels: A1B1, A1B2, A2B1, and A2B2 (Figure 1).

Second, the propensity scores can be estimated using multinomial logistic regression model, e.g., Group A1B1 as the reference group, all the other groups are compared with this reference group.

Third, various propensity score methods can be applied to estimate the group differences, which can be easily converted to the estimates of the main and interaction effects of two factors.

These propensity score applications include: subclassification (Rosenbaum & Rubin, 1984), optimal nonbipartite matching (Lu, Greevy, Xu, & Beck, 2011; Lu & Rosenbaum, 2004), inverse of propensity score weighting (Imbens, 2000), and using propensity score as covariate.

Monte Carlo Simulation

We conduct Monte Carlo Simulation for two scenarios: (1) semi-experiment, which concerns random assignment of Factor A while Factor B is not random assigned, and (2) non-experiment, i.e., neither Factor A nor Factor B concerns random assignment, and Factors A and B are correlated.

1. Semi-experiment

One example of this scenario is a design to study the main and interaction effects of new math curriculum (Factor A) and student socioeconomic status (SES) (Factor B): students are randomly assigned to treatment group (new math curriculum) and control group (business as usual), however, students are not randomly assigned to SES groups.

(1) Model to produce data

We adopted a revised model of Expression 9 in Schochet & Burghardt (2007, p. 101). The data are generated using the following equations:

$$B^* = W_1 + W_2 + u$$

$B = 1$ if $B^* > 0$ and $B = 0$, otherwise

$$Y = 100 + 5W_1 + 5W_2 + 5W_1^3 + 5W_2 \times A + 10A + 10B + 10A \times B + e \quad (1)$$

where W_1 and $W_2 \sim N(0,1)$, are two covariates. A is a dichotomous random variable indicating the status of Factor A (e.g., treatment and control). B is a dichotomous variable indicating the status of Factor B (e.g., dichotomous SES) and it is a function of two covariates. Y is the outcome. $u \sim N(0,1)$ and $e \sim N(0,100)$.

This revised model differs from Schochet & Burghardt's (2007) original model in that it includes three additional terms: (1) $10B$, indicating the effect of Factor B, which is closer to reality, e.g., the later achievement gaps between SES groups still exist even though their baseline conditions are equivalent, (2) $5W_1^3$, a higher-order term, and (3) $5W_2 \times A$, indicating that the effect of Factor A is associated with one baseline covariate. The conventional misspecified OLS regression model (e.g., omitting the higher-order or/and interaction term) would produce biased parameter estimates.

The main and interaction effects of A and B, i.e., the coefficients of A, B, and $A \times B$ are all 10[2]. We also allow the proportions of random assignment to Level 1 of Factor A to vary from 0.2 to 0.8 in a step of 0.1. Based on Model 1 we produce a sample with the total sample size N = 8,000 for each of seven proportion categories.

(2) Analysis of the main and interaction effects of two factors

We first analyze the full sample using an OLS model (Model 2). Note that Model 2 is a misspecified model, which simulates the reality that researchers might not know the correct model. These results serve as references for comparing with propensity score methods.

$$Y_i = \beta_0 + \beta_1(W_1)_i + \beta_2(W_2)_i + \beta_A A_i + \beta_B B_i + \beta_{A\times B} A_i B_i + e_i \quad (2)$$

Figure 4 shows the effects of two factors in terms of regression coefficients.

[2] The mean of W_2 in the full sample is 0, hence, the average effect of the term $5W_2 \times A$ is 0.

We then apply propensity score methods to analyze the data. Table 1 presents various propensity score applications. First, the multinomial logistic regression model is run to estimate the *generalized propensity score* (Imbens, 2000) using covariates W_1 and W_2. The *generalized propensity score* is the conditional probability of receiving treatment *t* given pre-treatment covariate *X*, i.e., $r(t, X) = pr(T = t \mid X = X)$. Thus, each participant will have four generalized propensity scores associated with four groups. The distributions of the generalized propensity scores of being in the *reference group* (or any other particular group) for the four groups are examined for overlap. The sample with common support can be obtained by excluding participants whose probabilities of being in the *reference group* are bigger than the minimum value of four maximum probabilities, or smaller than the maximum value of the four minimum probabilities (called "common support sample", Data 2 in Table 1).

This "common support sample" is further used for greedy matching by matching the *reference group* with all the other three groups, respectively, based on the probabilities of being in the *reference group*. We then have three matched samples and each matched sample contains one sample of *reference group*. We choose the common sample of being in the *reference group* among three matched sample. The final matching sample includes the common sample of being in the *reference group* and their matching sample in all other three groups (Data 3 in Table 1).

For the full sample and common support sample, we subclassify data into 9 subclasses based on the probabilities of being in the *reference group*, and obtain the weighted average parameter estimates across 9 subclasses[3]. We also use the inverse of the *generalized propensity score* as weight (Imbens, 2000) to analyze data with full sample and "common supported sample", respectively. The weight is $\frac{1}{r(T_i = t, X)}$, where T_i denotes the treatment that subject *i* actually received. Furthermore, using data with full sample we use the probabilities of being in the *reference group* as covariate to estimate the effects of two factors. All the estimates of the effects of two factors are conducted using OLS (Model 2) and ANOVA, respectively.

Finally, two independent binary logistic regression models are used to estimate the propensity scores for Factor A and Factor B, respectively. The common support sample (Data 4 in Table 1) is obtained by including overlap sample among four groups based on propensity scores for Factors A and B (Steps 1.1-1.3 in the procedure of "factorial propensity score matching method"). As illustrated in Figure 2, we apply 3×3 subclassification based on two propensity score functions and estimate the weighted treatment effect across 9 subclasses (Imai & van Dyk, 2004). We then apply the "factorial propensity score matching" procedure (Steps 2-4) to get final matching sample (Data 5). Both the OLS (Model 2) and ANOVA models are used for data analysis.

2. Non-experiment

(1) Model to produce data

In this scenario neither Factor A nor Factor B concerns random assignment and they are correlated. We also allow the proportion of Level 1 of Factor A and Factor B to vary. The data are generated using the following equations:

[3] We choose to use 9 subclasses here in order to make results comparable with 3×3 subclassification based on two propensity score functions.

$A^* = W_1 + W_2 + u_a$

$A = 1$ if $A^* < Q(A^*, P_a)$ and $A = 0$, otherwise

$B^* = W_2 + W_3 + u_b$

$B = 1$ if $B^* < Q(B^*, P_b)$ and $B = 0$, otherwise

$$Y = 100 + 5W_1 + 5W_2 + 5W_3 + 5W_2^3 + 5W_1 \times A + 5W_3 \times B + 10A + 10B + 10A \times B + e \quad (3)$$

where W_1, W_2, and $W_3 \sim N(0,1)$, are three covariates. A is a dichotomous variable indicating the status of Factor A and it is a function of covariates W_1 and W_2. B is a dichotomous variable indicating the status of Factor B and it is a function of covariates W_2 and W_3. Q is the quantile function of the normal distribution, i.e., the inverse of the cumulative distribution function. P_a and P_b are the proportions of Level 1 of Factor A and Factor B, respectively. Given the proportion of Level 1 of Factor A (or B), if A^* (or B^*) is smaller than the value that corresponds to the percentile, P_a (or P_b) of the normal distribution, then that individual is in Level 1. Y is the outcome. u_a, $u_b \sim N(0,1)$ and $e \sim N(0,100)$.

Model 3 is extended from Model 1. It included one more covariate. The main and interaction effects of A and B, i.e., the coefficients of A, B, and $A \times B$ are all 10. We allow the proportions of random assignment to Level 1 of Factor A and Factor B to vary: (P_a, P_b) = {(0.5, 0.5), (0.5, 0.3), (0.3, 0.3), (0.7, 0.3), (0.7, 0.7)}. Based on Model 3 we produce a sample with the total sample size N = 8,000 for each of five proportion categories.

(2) Analysis of the main and interaction effects of two factors

We conduct similar analyses with semi-experiment scenario. We first analyze the full sample using an OLS model (Model 4). Note that Model 4 is a misspecified model, which simulates the reality that researchers might not know the correct model. These results serve as references for comparing with propensity score methods.

$$Y_i = \beta_0 + \beta_1(W_1)_i + \beta_2(W_2)_i + \beta_3(W_3)_i + \beta_A A_i + \beta_B B_i + \beta_{A\times B} A_i B_i + e_i \quad (4)$$

We then apply propensity score methods to analyze the data. Table 2 presents various propensity score applications. Both the OLS (Model 4) and ANOVA models are used for data analysis.

We replicate 1000 times for producing and analyzing data. The estimates of bias and *MSE* (mean square error) of the parameters (θ) can be calculated as below.

$\hat{Bias}(\hat{\theta}) = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k - \theta$, where k = 1, 2,…, K. K is the number of valid replications.

$$\hat{MSE}(\hat{\theta}) = \frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}_k - \theta)^2 = \hat{\text{var}}(\hat{\theta}) + \left(\hat{Bias}(\hat{\theta})\right)^2$$

We estimate bias and *MSE* for β_A, β_B, and $\beta_{A\times B}$, respectively. We then calculate the sum of the absolute value of bias and the sum of *MSE* across these three parameters. Finally we calculate the percent reduction in summed bias and in summed *MSE* for propensity score methods as compared with the conventional OLS estimates.

Results and Conclusions:

Table 3 and Table 4 present the average percentages of the final analysis samples in the full sample for semi-experiment and non-experiment simulations, respectively. The common support samples for both propensity score models are pretty large. It is bigger than 85% in semi-experiment simulation and bigger than 68% in non-experiment simulation. However, it is not surprising that the matching samples are small because we used one-to-one matching and allow the sample allocation between two levels of Factor A and Factor B to vary. It is smaller than 38% in semi-experiment simulation and smaller than 20% in non-experiment simulation.

Table 5 and Table 6 present the results for percent reduction in the sum of the absolute value of bias and percent reduction in summed MSE (mean square error) among three parameters as compared with the conventional OLS estimate in semi-experiment simulation. For one multinomial propensity score model, application of common support sample has bigger bias and MSE reduction comparing with the full sample and OLS estimate has bigger bias and MSE reduction than ANOVA. These results are consistent with previous propensity scoring studies. In addition, including propensity score as covariate in the analysis has smaller effects in bias and MSE reduction than the other propensity score applications (e.g., subclassification and weighting). For two binary propensity score models, subclassification and factorial matching have similar good performance in MSE reduction as subclassification, weighing, and matching for one multinomial propensity score model when OLS regression is used, however, subclassification for two binary propensity score models has slightly smaller bias reduction than the others (e.g., 89% vs. 95%).

Table 7 and Table 8 present the results for percent reduction in the sum of the absolute value of bias and percent reduction in summed MSE (mean square error) among three parameters as compared with the conventional OLS estimate in non-experiment simulation. Similar with semi-experiment simulation, in general, OLS estimate has better performance than ANOVA in bias and MSE reduction. Thus, below we only concern OLS estimate. Regarding bias reduction, the three approaches better than the conventional OLS analysis are: (1) inverse of propensity score weighting based on one multinomial propensity score model (53-85%), (2) subclassification (44-83%) and (3) factorial matching (13-96%) based on two binary propensity score models. Regarding MSE reduction, the three approaches better than the conventional OLS analysis are: (1) subclassification (50-88%) and (2) factorial matching (5-87%) based on two binary propensity score models, and (3) inverse of propensity score weighting based on one multinomial propensity score model (4-79%). In general, subclassification based on two binary propensity score models has more stable good performance in bias and MSE reduction.

In sum, the simulation results from semi-experiment and non-experiment scenarios suggest three good propensity score applications in reducing bias and MSE of parameter estimates in analyzing two factors: (1) inverse of propensity score weighting based on one multinomial propensity score model, (2) subclassification and (3) factorial matching based on two binary propensity score models. Also note that the common support sample and covariate adjustment are preferred.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Hill, J. L., Brooks-Gunn, J. & Waldfogel, J. (2003). Sustained Effects of High Participation in an Early Intervention for Low-Birth-Weight Premature Infants. *Developmental Psychology.* 39(4):730–44.

Imai, K. & van Dyk, D. A. (2004). Causal Inference with General Treatment Regimes: Generalizing the Propensity Score. *Journal of the American Statistical Association. 99* (467): 854-866.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*. 87(3), 706–710.

Leslie, S., & Thiebaud, P. (2007). Using Propensity Scores to Adjust For Treatment Selection Bias. *SAS Global Forum 2007*. Retrieved March 8th, 2011 from http://www2.sas.com/proceedings/forum2007/184-2007.pdf

Lochman, J. E., Boxmeyer, C. L., Powell, N. P., Roth, D., & Windle, M. (2006). Masked intervention effects: Analytic methods for addressing low dosage of intervention. *New Directions for Evaluation, 110*, 19-32.

Lu, B., Greevy, R., Xu, X., & Beck, C. (2011). Optimal Nonbipartite Matching and Its Statistical Applications. *The American Statistician, 65* (1). 21-29.

Lu, B., & Rosenbaum, P. (2004). Optimal Pair Matching With Two Control Groups, *Journal of Computational and Graphical Statistics*, 13, 422–434.

Ming, K. & Rosenbaum P.R. (2001). A Note on Optimal Matching with Variable Controls Using the Assignment Algorithm. *Journal of Computational and Graphical Statistics*, *10* (3), 455-463.

Peck, L. R. (2003). Subgroup analysis in social experiments: Measuring program impacts based on post-treatment choice. *American Journal of Evaluation. 24* (2): 157-187.

Rosenbaum, P. R. (2002). *Observational Studies*, 2nd ed. New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983). 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*(387), 516-524.

Schochet, P. Z. & Burghardt, J. (2007). Using Propensity Scoring to Estimate Program-Related Subgroup Impacts in Experimental Program Evaluations."*Evaluation Review*, 31 (2), 95 – 120.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton-Mifflin.

Steiner, P. M., & Cook, D. (in press). Matching and propensity scores. In *The Oxford Handbook of Quantitative Methods*.

Appendix B. Tables and Figures

Not included in page count.

Tables

Table 1: Propensity Score Applications in Monte Carlo Simulation for Semi-Experiment

Propensity Score Model	Data	Application of Propensity Score	Covariate Adjustment[d]	Analysis Model
One multinomial propensity score model	1. Full sample	Subclassification[a]	No	ANOVA (subclassification)
			Yes	OLS (subclassification)
		Weighting[b]	No	ANOVA (weighting)
			Yes	OLS (weighting)
		As covariate[a]	No	ANCOVA (as covariate)
			Yes	OLS (as covariate)
	2. Common support sample (overlap among four groups)	Subclassification[a]	No	ANOVA (subclassification)
			Yes	OLS (subclassification)
		Weighting[b]	No	ANOVA (weighting)
			Yes	OLS (weighting)
	3. Overlap sample of matching one group with all other three groups using Data 2 above	Matching[a]	No	ANOVA
			Yes	OLS
Two binary propensity score models	4. Common support sample (overlap among four groups)	Subclassification[c]	No	ANOVA (subclassification)
			Yes	OLS (subclassification)
	5. Overlap sample of factorial matching using Data 4 above	Matching[c]	No	ANOVA (factorial matching)
			Yes	OLS (factorial matching)

Note:
[a]Subclassification, matching, and as covariate are all based on the probabilities of being in the *reference group*.
[b]Weighting is weighted by the inverse of the probabilities of being the group of treatment that participants actually received.
[c]Subclassification and matching are based on two propensity score functions.
[d]Covariate adjustment is to adjust for covariates W_1 and W_2.

Table 2: Propensity Score Applications in Monte Carlo Simulation for Non-Experiment

Propensity Score Model	Data	Application of Propensity Score	Covariate Adjustment[d]	Analysis Model
One multinomial propensity score model	1. Common support sample (overlap among four groups)	Subclassification[a]	No	ANOVA (subclassification)
			Yes	OLS (subclassification)
		Weighting[b]	No	ANOVA (weighting)
			Yes	OLS (weighting)
	2. Overlap sample of matching one group with all other three groups using Data 1 above	Matching[a]	No	ANOVA
			Yes	OLS
Two binary propensity score models	3. Common support sample (overlap among four groups)	Subclassification[c]	No	ANOVA (subclassification)
			Yes	OLS (subclassification)
	4. Overlap sample of factorial matching using Data 3 above	Matching[c]	No	ANOVA (factorial matching)
			Yes	OLS (factorial matching)

Note:
[a]Subclassification, matching, and as covariate are all based on the probabilities of being in the *reference group*.
[b]Weighting is weighted by the inverse of the probabilities of being the group of treatment that participants actually received.
[c]Subclassification and matching are based on two propensity score functions.
[d]Covariate adjustment is to adjust for covariates W_1, W_2, and W_3.

Table 3: Average Sample Sizes of Various Analyses in Semi-Experiment Simulation

Propensity Score Model	Data	Proportion of Level 1 of Factor A						
		0.2	0.3	0.4	0.5	0.6	0.7	0.8
One multinomial propensity score model	1. Full sample	100	100	100	100	100	100	100
	2. Common support sample (overlap among four groups)	86	88	90	90	90	90	90
	3. Overlap sample of matching one group with all other three groups using Data 2 above	8	15	22	28	27	22	15
Two binary propensity score models	4. Common support sample (overlap among four groups based on 2 propensity scores)	85	87	88	88	88	87	86
	5. Overlap sample of factorial matching using Data 4 above	15	23	30	37	30	23	15

Note: Entries are the average percentages of the final analysis samples in the full sample (N = 8000). 1000 replications.

Table 4: Average Sample Sizes of Various Analyses in Non-Experiment Simulation

Propensity Score Model	Data	Proportion of Level 1 of Factor A and Factor B				
		P_a =0.5 P_b =0.5	P_a =0.5 P_b =0.3	P_a =0.3 P_b =0.3	P_a =0.7 P_b =0.3	P_a =0.7 P_b =0.7
One multinomial propensity score model	1. Common support sample (overlap among four groups)	82	80	74	81	75
	2. Overlap sample of matching one group with all other three groups using Data 1 above	19	15	15	10	16
Two binary propensity score models	3. Common support sample (overlap among four groups based on 2 propensity scores)	77	73	69	69	68
	4. Overlap sample of factorial matching using Data 3 above	15	13	12	9	12

Note: Entries are the average percentages of the final analysis samples in the full sample (N = 8000). 1000 replications.

Table 5: Percent Bias Reduction in Semi-Experiment Simulation

Propensity Score Model	Data	Analysis Model	Proportion of Level 1 of Factor A							Average
			0.2	0.3	0.4	0.5	0.6	0.7	0.8	
One multinomial propensity score model	1. Full sample	ANOVA (subclassification)	75	74	75	73	75	75	75	75
		OLS (subclassification)	93	91	91	91	94	91	94	92
		ANOVA (weighting)	48	54	56	56	59	61	63	57
		OLS (weighting)	84	86	87	87	88	88	89	87
		ANCOVA (including propensity score as covariate)	28	26	26	25	24	23	22	25
		OLS (including propensity score as covariate)	25	24	23	22	21	20	20	22
	2. Common support sample (overlap among four groups)	ANOVA (subclassification)	94	92	90	88	88	86	85	89
		OLS (subclassification)	97	96	95	95	95	95	95	95
		ANOVA (weighting)	84	87	92	90	88	86	83	87
		OLS (weighting)	98	98	97	96	96	96	96	97
	3. Overlap sample of matching one group with all other three groups using Data 2 above	ANOVA	92	95	96	96	96	97	97	96
		OLS	92	94	95	96	97	98	98	96
Two binary propensity score models	4. Common support sample (overlap among four groups based on 2 propensity scores)	ANOVA (subclassification)	69	68	67	68	69	71	73	69
		OLS (subclassification)	90	90	89	89	89	89	90	89
	5. Overlap sample of factorial matching using Data 4 above	ANOVA (factorial matching)	100	99	99	99	99	100	99	99
		OLS (factorial matching)	99	99	100	100	99	100	99	99

Note: Entries are percent reduction in the sum of the absolute value of bias among three parameters as compared with the conventional OLS estimate. 1000 replications.

Table 6: Percent Reduction in Summed MSE (Mean Square Error) in Semi-Experiment Simulation

Propensity Score Model	Data	Analysis Model	Proportion of Level 1 of Factor A							Average
			0.2	0.3	0.4	0.5	0.6	0.7	0.8	
One multinomial propensity score model	1. Full sample	ANOVA (subclassification)	56	64	69	69	70	66	62	65
		OLS (subclassification)	79	85	88	88	89	88	86	86
		ANOVA (weighting)	9	21	34	37	36	41	26	29
		OLS (weighting)	85	88	89	90	89	89	84	88
		ANCOVA (including propensity score as covariate)	29	32	35	37	39	39	40	36
		OLS (including propensity score as covariate)	30	33	35	37	38	38	38	36
	2. Common support sample (overlap among four groups)	ANOVA (subclassification)	89	90	90	89	89	87	85	88
		OLS (subclassification)	92	94	95	95	95	95	93	94
		ANOVA (weighting)	74	80	82	83	83	80	71	79
		OLS (weighting)	91	94	94	94	94	93	88	93
	3. Overlap sample of matching one group with all other three groups using Data 2 above	ANOVA	78	90	94	95	96	95	93	92
		OLS	81	91	94	96	96	96	94	93
Two binary propensity score models	4. Common support sample (overlap among four groups based on 2 propensity scores)	ANOVA (subclassification)	76	76	77	78	80	82	84	79
		OLS (subclassification)	95	96	96	96	96	96	96	96
	5. Overlap sample of factorial matching using Data 4 above	ANOVA (factorial matching)	88	93	95	97	96	95	93	94
		OLS (factorial matching)	92	95	96	97	97	96	95	95

Note: Entries are percent reduction in summed MSE among three parameters as compared with the conventional OLS estimate. 1000 replications.

Table 7: Percent Bias Reduction in Non-Experiment Simulation

Propensity Score Model	Data	Analysis Model	Proportion of Level 1 of Factor A and Factor B				
			P_a =0.5 P_b =0.5	P_a =0.5 P_b =0.3	P_a =0.3 P_b =0.3	P_a =0.7 P_b =0.3	P_a =0.7 P_b =0.7
One multinomial propensity score model	1. Common support sample (overlap among four groups)	ANOVA (subclassification)	-51	-199	-90	-59	13
		OLS (subclassification)	-94	-242	-142	-40	4
		ANOVA (weighting)	85	44	68	71	86
		OLS (weighting)	85	53	61	78	85
	2. Overlap sample of matching one group with all other three groups using Data 1 above	ANOVA	-54	-246	-128	-56	51
		OLS	-110	-307	-217	-37	14
Two binary propensity score models	3. Common support sample (overlap among four groups based on 2 propensity scores)	ANOVA (subclassification)	3	-76	-39	48	56
		OLS (subclassification)	75	44	50	76	83
	4. Overlap sample of factorial matching using Data 3 above	ANOVA (factorial matching)	98	13	49	30	70
		OLS (factorial matching)	96	13	43	30	72

Note: Entries are percent reduction in the sum of the absolute value of bias among three parameters as compared with the conventional OLS estimate. 1000 replications.

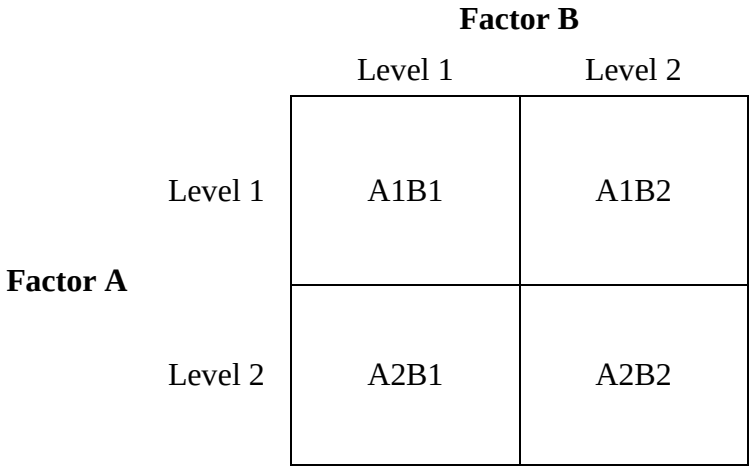
Table 8: Percent Reduction in Summed MSE (Mean Square Error) in Non-Experiment Simulation

Propensity Score Model	Data	Analysis Model	Proportion of Level 1 of Factor A and Factor B				
			P_a =0.5 P_b =0.5	P_a =0.5 P_b =0.3	P_a =0.3 P_b =0.3	P_a =0.7 P_b =0.3	P_a =0.7 P_b =0.7
One multinomial propensity score model	1. Common support sample (overlap among four groups)	ANOVA (subclassification)	-167	-407	-159	-280	49
		OLS (subclassification)	-301	-431	-302	-152	32
		ANOVA (weighting)	-243	-208	-177	-38	52
		OLS (weighting)	4	6	10	48	79
	2. Overlap sample of matching one group with all other three groups using Data 1 above	ANOVA	-138	-497	-269	-264	78
		OLS	-350	-605	-580	-124	46
Two binary propensity score models	3. Common support sample (overlap among four groups based on 2 propensity scores)	ANOVA (subclassification)	-21	-83	-87	42	80
		OLS (subclassification)	74	50	54	72	88
	4. Overlap sample of factorial matching using Data 3 above	ANOVA (factorial matching)	64	-5	24	11	84
		OLS (factorial matching)	70	5	31	15	87

Note: Entries are percent reduction in summed MSE among three parameters as compared with the conventional OLS estimate. 1000 replications.

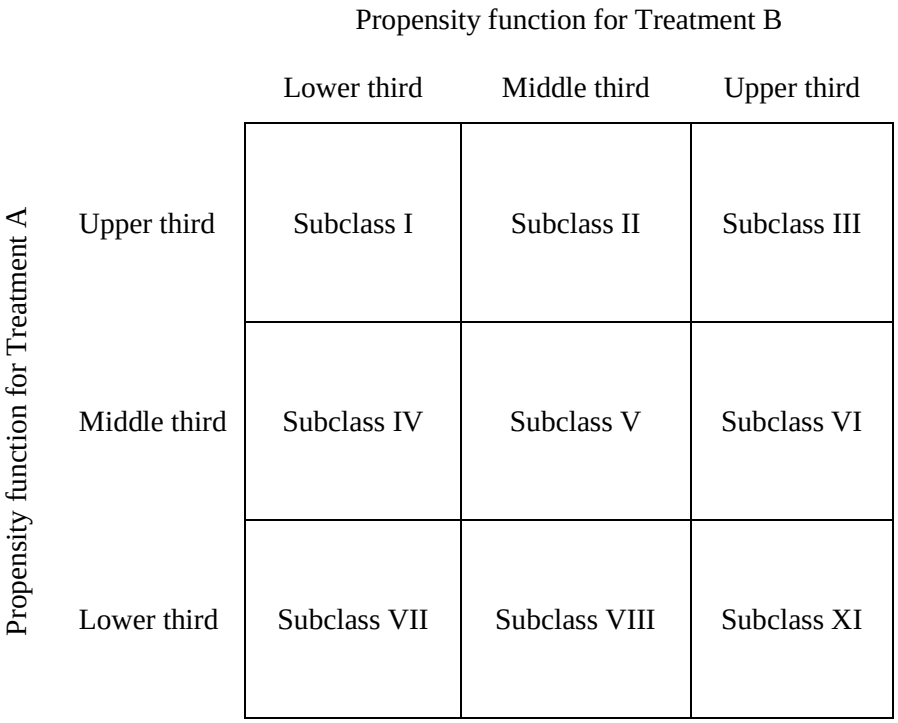
Figures

Figure 1. 2×2 Factorial Design



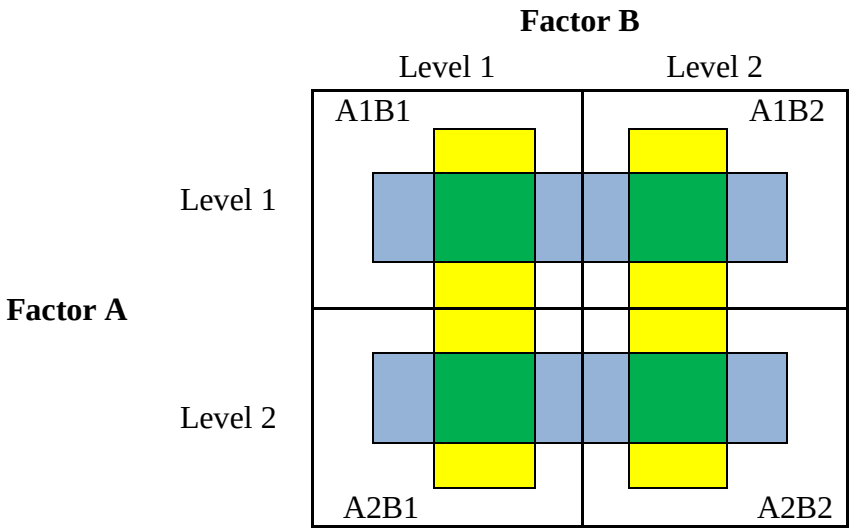
Note: Adapted From Figure 8.1 by Shadish, Cook, & Campbell (2002, p.264)

Figure 2. 3 × 3 Subclassification Based on Two Propensity Score Functions (Imai & van Dyk, 2004)



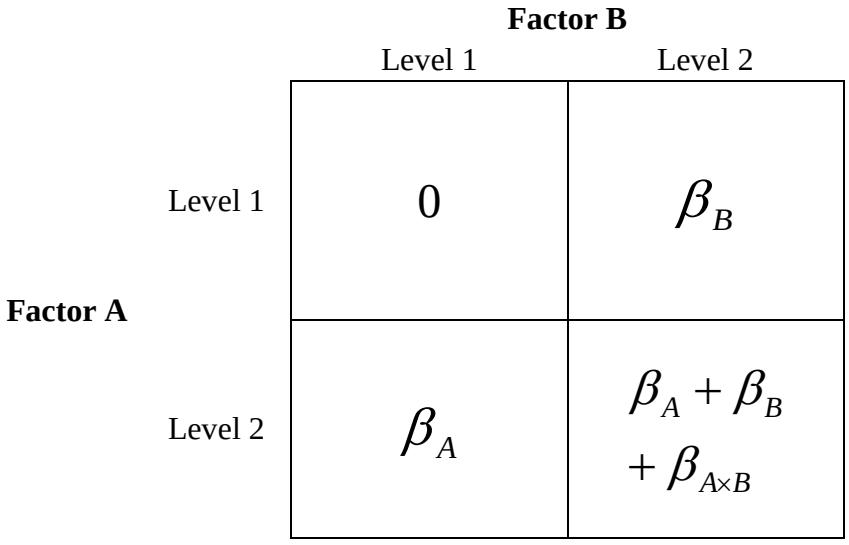
Note: Adapted from Figure 4 by Imai & van Dyk (2004, p.861). Data are subclassified into three subclasses (lower third, middle third, and upper third) based on each of two propensity score functions, respectively. Each cell of the 3 × 3 table represents a subclass based on two propensity score functions jointly.

Figure 3. Factorial Propensity Score Matching



Note: Yellow and green areas represent “matched on A”; blue and green areas represent “matched on B”. The four green areas represent the common participants in two matched datasets (Data 5 in Table 1).

Figure 4. Effects of Two Factors in Terms of Regression Coefficients



Note: The cell of Level 1 of Factor A and Level 1 of Factor B serving as the reference group.