Abstract Title Page

Title:

Empirical Benchmarks of Hidden Bias in Educational Research: Implication for Assessing How Well Propensity Score Methods Approximate Experiments and Conducting Sensitivity Analysis

Authors and Affiliations:

Nianbo Dong
University of Missouri

Mark Lipsey
Vanderbilt University

Abstract Body

Background / Context:

When randomized control trials (RCT) are not feasible, researchers seek other methods to make causal inference, e.g., propensity score methods (Rosenbaum & Rubin, 1983). One of the underlined assumptions for the propensity score methods to obtain unbiased treatment effect estimates is the ignorability assumption, that is, conditional on the propensity score, treatment assignment is independent of the outcome. However, this assumption is hard to empirically test. In other words, researchers who used propensity score methods did not know how well the ignorability assumption can be met in their research. Sensitivity analysis, e.g., Rosenbaum's (2002) Gamma parameter based on Wilcox rank statistics, and other statistics based on regression (Frank, 2000; Hong & Raudenbush, 2006; Lin, Psaty, & Kronmal, 1998; Pan & Frank, 2003), could be conducted to assess the sensitivity of a statistical conclusion when the ignorability assumption is not met (i.e., assuming certain magnitude of hidden bias due to unmeasured confounders), however it usually lacks empirical evidence regarding how large the hidden bias could be reasonable in educational studies given that the demographic information and pretest are available.

Using the results from the experiments as benchmark, the within-study comparison designs allow the researchers to create another comparison group based on quasi-experimental designs to estimate the intervention effects and empirically assess how well this particular quasi-experimental design under certain conditions can approximate experiments, and researchers have drawn different conclusions regarding if quasi-experiments can replicate experiments (e.g., Fraker & Maynard, 1987; Heckman, Hotz, & Dabos, 1987; Michalopoulos, Bloom, & Hill, 2004; Wilde & Hollister, 2007). In particular, Cook and colleagues (e.g., Cook, Shadish, & Wong, 2008; Cook, Steiner, & Pohl, 2010; Pohl, Steiner, Eisermann, Shadish, Clark, & Steiner, 2008; Steiner, Cook, Shadish, & Clark, 2010; Wong, Hallberg, & Cook, 2013) have used within-study comparisons to identify under what conditions (e.g., covariates selection, matching within or between locations/clusters, etc.) the quasi-experiment can replicate experiments.

Some useful suggestions about constructing a good comparison group have been made, e.g., using local matching and including pretests in matching (Cook, Shadish, & Wong, 2008; Michalopoulos, Bloom, & Hill, 2004; Steiner, Cook, Shadish, & Clark, 2010). In particular, Wong, Hallberg, & Cook (2013) examined the relative importance of focal and local matching and concluded that intact school matching within districts can replicate experimental estimates. Although advances have been made in this area, as Cook (2012) suggested, more within-study comparisons are needed to assess the robustness of ability that well designed and implemented quasi-experiments replicate experiments across different populations, settings, and times, etc. In addition, the within-study comparison designs provide a useful approach to empirically estimating the hidden bias due to unmeasured confounders for the propensity score applications under certain conditions.

Purpose / Objective / Research Question / Focus of Study:

The purpose of this study is to use within-study comparisons to assess how well propensity score methods can approximate experiments under various conditions. In particular, we test three ways of constructing comparison groups: (1) using the sample from the states that are different from

the original experiments, with pretest and demographic information, (2) using the sample from the same state and districts (local matching) with the experiments, with demographic information only, and (3) same as (2) but with pretest as well. Propensity score methods (optimal matching, propensity score weighting, and stratification) are used to estimate the treatment effects for three ways of constructing comparison groups, which are compared with the benchmark from the experiment to assess estimate bias.

Significance / Novelty of study:

This study will contribute to the literature by providing empirical evidence about how well propensity score methods can approximate experiments under various conditions. In addition, the bias estimated from using propensity score methods is hidden bias due to unmeasured confounders, which can provide reference information about the magnitude of hidden bias for sensitivity analysis to assess robustness of the propensity score estimates under different conditions.

Research Design:

Data
This study uses data from four IES funded projects, among which three are large scaled experiments: (1) "Scaling up TRIAD: Teaching Early Mathematics for Understanding with Trajectories and Technologies" (Clements & Sarama, 2006), (2) "Evaluating the Effectiveness of Tennessee's Voluntary Pre-Kindergarten Program" (Lipsey, et al., 2011 & 2013), and (3) "Experimental Evaluation of the Tools of the Mind Pre-K Curriculum" (Wilson & Farran, 2013), and one is a measurement study, "Learning-Related Cognitive Self-Regulation School Readiness Measures for Preschool Children Study" (Lipsey & Meador, 2013).

The "Scaling up TRIAD" study was a project that was to evaluate the effects of preschool mathematics intervention across three sites (Buffalo, NY; Boston, MA; Nashville, TN). A cluster randomized control trial in which schools were randomly assigned to the treatment and control conditions was conducted for each site. The NY site had a sample of 25 schools and 946 students, the MA site had a sample of 18 schools and 359 students, and the TN site had a sample of 16 schools and 409 students (Hofer, Lipsey, Dong, & Farran, 2013). The common variables collected across three sites included: (1) pre- and post-test of outcome: Research-based Elementary Math Assessment (REMA), a proximal measure of children's early math skills (Clements, Sarama, & Liu, 2008), and (2) child demographic information (race, gender, age, language spoken at home, and mother's highest education). In addition, the TN site collected the pre- and post-test outcomes on Woodcock Johnson III Achievement Battery (Woodcock, McGrew, and Mather, 2001) that included Applied Problems, Quantitative Concepts, and Letter-Word Identification, etc. Table 1 lists the descriptive statistics of covariates by site and by treatment conditions.

The Tennessee PreK Evaluation was to evaluate the effectiveness of the Tennessee Voluntary Pre-K program (Lipsey, et al., 2011 & 2013). It consists of a blocked individual random assignment design and a regression discontinuity design. The total sample included 59 schools and more than 2000 students. The pre- and post-test on Woodcock Johnson measures and the child demographic information were collected.

The Tools of the Mind study applied a cluster randomized design to evaluate the effectiveness of the Tools of the Mind Curriculum (Wilson & Farran, 2013). Sixty prekindergarten classrooms in Tennessee and North Carolina were randomly assigned to the treatment (Tools classroom) and control conditions. More than 800 children were collected data on the pre- and post-test on Woodcock Johnson measures and the child demographic information.

The self-regulation study was a measurement project aiming to identify a set of direct assessment measures for learning-related cognitive self-regulation school readiness measures that could predict academic achievement (Lipsey & Meador, 2013). More than 500 pre-k children in 38 schools/centers in Tennessee were collected data on self-regulation measures, Woodcock Johnson measures, and the child demographic information.

Analytic Plan

The treatment effect estimated from the cluster randomized control trial at the Tennessee site in the "Scaling up TRIAD" project serve as the benchmark. In this well implemented experimental study, the "average effect of the treatment on the treated" (ATT) and the "average treatment effect" (ATE) on all samples (Imai, King, & Stuart, 2008; Imbens, 2004; Mccaffrey, Ridgeway, & Andmorral, 2004; Ridgeway et al., 2012) should be identical. We target at the population that the sample ($N_t = 211$) in the treatment group at the TN site represented. The comparison groups are constructed to serve as the counterfactuals of the treatment group. Hence, we focus on ATT, which is estimated from the total sample of the treatment group at TN site and the comparisons groups constructed using different samples and methods.

We construct comparison groups using three ways: (1) using the samples from different states: the control groups from the MA and NY sites in the same project ("Scaling up TRIAD") and from North Carolina in the different project ("Tools of the Mind"), (2) using the samples from the same state (TN): control sample from the different project ("Tennessee PreK Evaluation"), and the whole sample from the measurement study ("self-regulation study") with demographic information only, and (3) same as (2) but with pretest as well.

The propensity scores are estimated using the combined sample from the treated sample in the Tennessee site in the "Scaling up TRIAD" project and one comparison group. Three types of propensity score methods used to estimate the "average effect of the treatment on the treated" (ATT) include: (1) One-to-one optimal matching (Ming & Rosenbaum, 2001), i.e., matching the treated sample at the TN site in the "Scaling up TRIAD" project with the sample from different pools of comparison groups listed above, (2) Weighting by the odds of the propensity score, i.e., the sample in the treatment group has a weight of 1, and the sample in the comparison group has a weight of $\frac{\hat{e}_i}{1-\hat{e}_i}$, where $\hat{e}_i$ is the estimated propensity score (Hirano, Imbens, & Ridder, 2003), and (3) Stratification, i.e., the sample is stratified to 5 groups based on the estimated propensity score, and the ATT is estimated by the average treatment effects across 5 strata weighted by the proportion of the sample size for the treatment group in each stratum (Rosenbaum & Rubin, 1984).

The point estimates and their 95% confidence intervals using propensity score methods on different samples are compared with the benchmark (point estimate and its 95% confidence interval of the math curriculum treatment effect in Tennessee). The estimate bias is calculated by

the difference in point estimates between the propensity score methods and the benchmark, however, the estimation errors should be considered using the 95% confidence intervals.

Results and Conclusions:

The analysis is undergoing. We report partial results here, in which the counterfactuals were constructed using the samples in the control groups from the MA and NY sites in the "Scaling up TRIAD" project. Table 1 reports the descriptive statistics of covariates and covariate balance checking between the treatment and control groups for the cluster randomized trials in three sites (TN, NY, & MA) in the "Scaling up TRIAD" project. For the TN site, pretest and other eight covariates except mother's highest education are balanced with the standardized mean difference smaller than 0.25 between the treatment and control groups. The 211 children in the treatment group in TN serve as the focal treated sample that we would like to estimate the treatment effect. The 286 children in the control group in NY and 92 children in the control group serve as the pool for constructing the counterfactuals of treated sample.

Table 2 presents the covariate balancing checking for the matched samples using NY and MA control groups (Column 2), and using NY, MA, and TN control groups (Column 3) based on 1-to-1 optimal matching. The two matched samples had covariates close to the focal treated sample (Column 1).

Table 3 presents the ATT estimates in effect size and their 95% confidence intervals for the cluster randomized trials in three sites (TN-benchmark, NY, & MA), and effect size estimates, their 95% confidence intervals, bias, and the percentage of bias (100*(bias/benchmark)) for the ATT estimates using the propensity score methods for different comparison samples. The effect size benchmark for the TN treated sample is 0.63 with a 95% confidence interval of (0.38, 0.88). The ATT estimate (0.61) from the experiment at NY is similar with TN, while the ATT estimate (0.29) from the experiment at MA is quite different from TN but not statistically different at an alpha of 0.05. The bias and percentage of bias for the propensity score estimates range from -0.10 to -0.21 and from -15.3% to -34.0%, and they are not statistically significant at an alpha of 0.05. The propensity score estimate using the MA control sample produced the biggest bias (-0.21) and the propensity score estimate using the NY control sample produced less bias (-0.11). The different propensity score methods (optimal matching, weighting, and stratification) using the same sample produced very consistent estimates.

Figure 1 illustrates the effect sizes and their 95% confidence intervals of various ATT estimates using data from Table 3. It is very clear that all the 95% confidence intervals of the ATT estimates using the propensity score methods cover the point estimates of the benchmark.

In sum, constructing the comparison groups using the cross-state sample produced statistically non-significant but sizable bias. We are working on constructing the comparison groups using local matching and expect to have smaller bias. However, "how close is close enough" (Wilde & Hollister, 2007) still remains questions and more studies about the criteria for assessing the quality of propensity score methods in replicating experiments are needed. Nevertheless, these bias estimates will provide reference values used for sensitivity analysis to assess the robustness to violation of the independence assumption in applying propensity score methods to educational research.

Appendices

Appendix A. References

Clements, D. & Sarama, J. (June 2006). *Scaling Up TRIAD: Teaching Early Mathematics for Understanding with Trajectories and Technology*, proposal funded by the Institute for Education Sciences.

Clements, D., Sarama, J., & Liu, X. (July 2008). Development of a measure of early mathematics achievement using the Rasch model: the Research-Based Early Math Assessment. *Educational Psychology, 28* (4), 457 - 482.

Cook, T. D. (2012). *Introduction: Theory of Within-Study Comparison Design.* Workshop presentation at the 2012 IES/NCER Summer Research Training Institute: Design, Implementation, and Analysis of Within-Study Comparisons. Retrieved at http://www.ipr.northwestern.edu/workshops/past-workshops/design-implementation-and-analysis-of-within-study-comparisons/2012/docs/WSC%20workshop%20Day%201%20ppt.pdf

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies often produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724-750.

Cook, T. D., Steiner, P., & Pohl, S. (2010). Assessing how bias reduction is influenced by covariate choice, unreliability and data analytic mode: An analysis of different kinds of within-study comparisons in different substantive domains. *Multivariate Behavioral Research 44*(6): 828-47.

Fraker, T. & Maynard, R. A. (1987). Evaluating Comparison Group Designs with Employment-Related Programs. *Journal of Human Resources*, 22: 194–227.

Frank, K. A. (2000). The impact of a confounding variable on a regression coefficient. *Sociological Methods and Research, 29*(2), 147-194.

Heckman, J. J., Hotz, V. J., & Dabos, M. (1987). Do we need experimental data to evaluate the impact of interventions?, *Evaluation Review, 11* (4), 395-427.

Hofer, K. G., Lipsey, M. W., Dong, N., & Farran, D. C. (2013). *Results of the Early Math Project – Scale-Up Cross-Site Results (Working Paper)*. Nashville, TN: Vanderbilt University, Peabody Research Institute. Available at http://peabody.vanderbilt.edu/research/pri/Early_Math_Cross-Site_Technical_Report_Working_Paper.pdf

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71*, 1161–1189.

Hong, G. & Raudenbush, S. W. (2006). Evaluating Kindergarten Retention Policy: A Case Study of Causal Inference for Multilevel Observational Data. *Journal of the American Statistical Association. 101* (475), 901-910.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists in causal inference. Journal of the Royal Statistical Society, Series A, 171, 481–502.

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

Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998). Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies. *Biometrics*, 54, 948–963.

Lipsey, M. W., Farran, D. C., Bilbrey, C., Hofer, K. G., & Dong, N. (2011). *Initial results of the evaluation of the Tennessee voluntary pre-k program*. Working Paper. Peabody Research Institute, Vanderbilt University.

Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., & Bilbrey, C. (2013). *Evaluation of the Tennessee Voluntary Prekindergarten Program: Kindergarten and First Grade Follow-Up Results from the Randomized Control Design (Research Report).* Nashville, TN: Vanderbilt University, Peabody Research Institute.

Lipsey, M. W. & Meador, D. N. (2013). *Learning-Related Cognitive Self-Regulation School Readiness Measures for Preschool Children*. IES PI Meeting. Washington, DC.

Michalopolous, C., Bloom, H., & Hill, C. J. (2004). Can propensity score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? The Review of Economics and Statistics, 86, 156–179.

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.

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

Pan, W. & Frank, K. A. (2003). A probability index of the robustness of a causal inference. *Journal of Educational and Behavioral Statistics, 28* (4), 315-337.

Pohl, S., Steiner, P. Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis, 31*(4): 463-79.

Ridgeway, G. McCaffey, D. F., Morral, A., Burgette, L., & Griffin, B. A. (2012). Toolkit for Weighting and Analysis of Nonequivalent Groups: A tutorial for the twang package. Retrieved on September 6, 2012 on http://cran.r-project.org/web/packages/twang/vignettes/twang.pdf

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies 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.

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

Steiner, P.M., Cook, T.D., Shadish, W.R., & Clark, M.H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*(3), 250-267.

Shadish, W.R., Clark, M.H., & Steiner, P.M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment (with comments by Little/Long/Lin, Hill, and Rubin, and a rejoinder). *Journal of the American Statistical Association, 103,* 1334-1356.

Wong, V. C., Hallberg, K., & Cook, T. D. (2013). Intact school matching in education: Exploring the relative importance of focal and local matching. Paper presented at the SREE Spring 2013 Conference. Retrieved on May 1st, 2013 from https://www.sree.org/conferences/2013s/program/downloads/abstracts/842.pdf

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III. Itasca, IL: Riverside Publishing.

Wilde, E.T., & Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis and Management* 26*:*455-77.

Wilson, S. J. & Farran, D. C. (2013). Experimental Evaluation of the Tools of the Mind Preschool Curriculum. Paper presented at the SREE Spring 2012 Conference. Retrieved on May 1st, 2012 from https://www.sree.org/conferences/2012s/program/downloads/abstracts/437_1.pdf

Appendix B. Tables and Figures

Table 1: Covariate Balance Checking between the Treatment and Control Groups by Site

	TN			NY			MA		
Variable	Treatment	Control	Effect Size (T-C)	Treatment	Control	Effect Size (T-C)	Treatment	Control	Effect Size (T-C)
Pretest	38.13 (6.01)	37.64 (5.64)	0.09	38.10 (5.89)	38.72 (5.43)	-0.11	39.21 (6.23)	39.85 (6.56)	-0.10
Age (month)	60.36 (3.96)	60.71 (3.68)	-0.09	58.6 (3.58)	58.77 (3.79)	-0.04	62.61 (4.00)	63.01 (3.91)	-0.10
Interval between pre- and post-test (month)	7.34 (0.48)	7.22 (0.54)	0.24	7.99 (0.50)	7.09 (0.61)	1.68	7.71 (0.55)	7.16 (0.98)	0.79
Test lag of pretest from school start date	1.03 (0.47)	1.08 (0.44)	-0.11	0.67 (0.50)	1.38 (0.49)	-1.42	0.78 (0.44)	1.12 (0.33)	-0.82
Black	0.81 (0.40)	0.72 (0.45)	0.20	0.66 (0.47)	0.55 (0.50)	0.22	0.30 (0.46)	0.30 (0.46)	0.00
White	0.06 (0.23)	0.12 (0.32)	-0.21	0.22 (0.42)	0.19 (0.39)	0.09	0.11 (0.32)	0.12 (0.33)	-0.02
Hispanic	0.08 (0.27)	0.13 (0.34)	-0.18	0.08 (0.27)	0.19 (0.39)	-0.35	0.48 (0.50)	0.49 (0.50)	-0.01
ELL	0.09 (0.28)	0.14 (0.34)	-0.16	0.03 (0.16)	0.16 (0.37)	-0.56	0.42 (0.49)	0.45 (0.50)	-0.06
Male	0.46 (0.50)	0.44 (0.50)	0.03	0.50 (0.50)	0.49 (0.50)	0.01	0.46 (0.50)	0.51 (0.50)	-0.11
Mother's highest Education	1.48 (0.91)	1.17 (0.91)	0.34	1.56 (0.91)	1.40 (0.96)	0.17	1.58 (0.92)	1.55 (1.01)	0.03
N	211	198		660	286		267	92	

Note: Entries are means and standard deviations (in parenthesis).

Table 2: Covariate balance checking for the matched samples based on 1-to-1 optimal matching

Variable	(1)Treatment (TN)	(2)Control (NY+MA)	(3)Control (NY+MA+TN)	Effect Size (1-2)	Effect Size (1-3)
Pretest	38.13 (6.01)	38.87 (5.43)	37.80 (5.35)	-0.13	0.06
Age (month)	60.36 (3.96)	59.96 (4.39)	60.19 (4.39)	0.10	0.04
Interval between pre- and post-test (month)	7.34 (0.48)	7.19 (0.82)	7.37 (0.84)	0.23	-0.03
Test lag of pretest from school start date	1.03 (0.47)	1.18 (0.39)	1.08 (0.44)	-0.35	-0.10
Black	0.81 (0.40)	0.74 (0.44)	0.81 (0.40)	0.15	0.00
White	0.06 (0.23)	0.07 (0.26)	0.04 (0.19)	-0.06	0.09
Hispanic	0.08 (0.27)	0.09 (0.28)	0.08 (0.27)	-0.03	-0.02
ELL	0.09 (0.28)	0.12 (0.32)	0.11 (0.31)	-0.11	-0.08
Male	0.46 (0.50)	0.45 (0.50)	0.46 (0.50)	0.03	0.00
Mother's highest Education	1.48 (0.91)	1.48 (0.95)	1.44 (0.99)	-0.01	0.04
N	211	211	211		

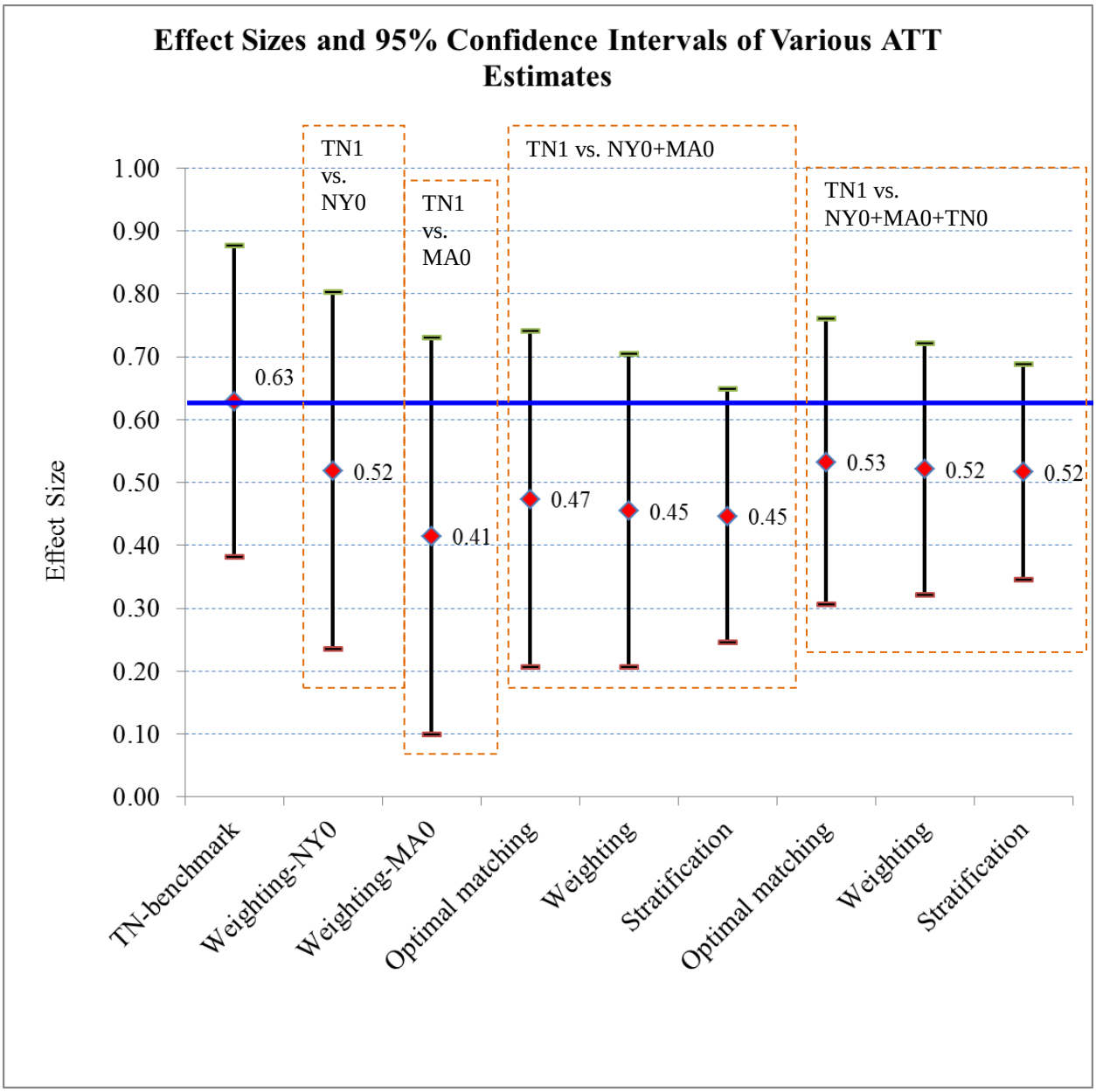
Note: Entries are means and standard deviations (in parenthesis).
(1)Treatment (TN) is the treatment group in TN, (2)Control (NY+MA) is the matched sample from the control groups in NY and MA based on 1-to-1 optimal matching, (3)Control (NY+MA+TN) is the matched sample from the control groups in NY, MA, and TN based on 1-to-1 optimal matching.

Table 3: Effect Sizes, 95% Confidence Intervals, Bias, and Percentage of Bias of Various Average Treatment Effect on Treated (ATT) Estimates

Sample	Analytic Method	Effect Size	95% CI		Bias[a]	Percentage of Bias[b]
			Lower	Upper		
TN (benchmark)	HLM	0.63	0.38	0.88	NA	NA
NY	HLM	0.61	0.39	0.82	NA	NA
MA	HLM	0.29	-0.01	0.60	NA	NA
TN1 vs. NY0	Weighting	0.52	0.23	0.80	-0.11	-17.5
TN1 vs. MA0	Weighting	0.41	0.10	0.73	-0.21	-34.0
TN1 vs. NY0+MA0	Optimal matching	0.47	0.21	0.74	-0.16	-24.7
TN1 vs. NY0+MA0	Weighting	0.45	0.21	0.70	-0.17	-27.6
TN1 vs. NY0+MA0	Stratification	0.45	0.24	0.65	-0.18	-29.0
TN1 vs. NY0+MA0+TN0	Optimal matching	0.53	0.31	0.76	-0.10	-15.3
TN1 vs. NY0+MA0+TN0	Weighting	0.52	0.32	0.72	-0.11	-17.1
TN1 vs. NY0+MA0+TN0	Stratification	0.52	0.34	0.69	-0.11	-17.8

Note: Entries are means and standard deviations (in parenthesis).
[a]Bias is calculated by the difference between the effect sizes estimated by the propensity score methods and the benchmark (0.63). [b]Percentage of Bias is calculated by 100*(Bias/0.63).

Figure 1. Effect Sizes and 95% Confidence Intervals of Various Average Treatment Effect on Treated (ATT) Estimates



Note: Blue line represents the impact benchmark from the TN experiment.