WCER Working Paper No. 2010-7
May 2010

Hand Matching vs. Propensity Score Matching: An Empirical Comparison of Results From a Quasi-Experiment

Natalie A. Tran
School of Natural Sciences and Mathematics
California State University–Bakersfield
ntran6@csub.edu

Alan B. Nathan
Department of Educational Leadership & Policy Analysis
University of Wisconsin–Madison
anathan2@wisc.edu

Mitchell J. Nathan
Department of Educational Psychology
Wisconsin Center for Education Research
University of Wisconsin–Madison
mnathan@wisc.edu

WCER
Wisconsin Center for Education Research
School of Education • University of Wisconsin–Madison • http://www.wcer.wisc.edu/

WCER working papers are available on the Internet at http://www.wcer.wisc.edu/publications/workingPapers/index.php. Recommended citation:

> Tran, N. A., Nathan, A. B., & Nathan, M. J. (2010). *Hand matching vs. propensity score matching: An empirical comparison of results from a quasi-experiment* (WCER Working Paper No. 2010-7). Retrieved from University of Wisconsin–Madison, Wisconsin Center for Education Research website: http://www.wcer.wisc.edu/publications/workingPapers/papers.php

The research reported in this paper was supported by a grant from the National Science Foundation (EEC-0648267) titled "Aligning Educational Experiences with Ways of Knowing Engineering" (AWAKEN) and by the Wisconsin Center for Education Research, School of Education, University of Wisconsin–Madison. Any opinions, findings, or conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies, WCER, or cooperating institutions.

Hand Matching vs. Propensity Score Matching: An Empirical Comparison of Results From a Quasi-Experiment

Natalie A. Tran, Alan B. Nathan, and Mitchell J. Nathan

Many education research studies are observational in nature. In an observational study, potential participants decide of their own volition whether to participate. The reasons behind these decisions can be multifaceted and are generally not documented. An example of such participation is course enrollment. Because participation is optional, it is difficult to make causal inferences based on participation. This is the well-known problem of selection bias.

Proponents of quasi-experimental methods in education have advocated *propensity score matching* (PSM) as a viable method for approximating the randomized assignment process that takes place in a true experiment, thereby reducing the effects of self-selection (Shadish, Cook, & Campbell, 2002; Rosenbaum & Rubin 1984). The propensity score can be thought of as the likelihood that participants will choose a treatment, based on their observed covariates. Participants with similar propensity scores who vary from each other only in assignment to treatment can therefore be regarded as the treatment and comparison cases. Differences in outcome can be attributed to program participation—either in the treatment or comparison group—rather than differences in the other covariates.

Randomized controlled trials have been identified as the "gold standard" of behavioral research (U.S. Department of Education, 2003). However, it is not always feasible to conduct experimental studies considering the limited amount of resources available and ethical dilemmas associated with program assignments. Therefore, we turn to other methodological approaches to advance research in the social sciences. While PSM has not yet achieved parity with randomized controlled trials as an evaluative approach (Cook, Shadish, & Wong, 2008; Wilde & Hollister, 2007), it is valuable to know how PSM compares with other methods of addressing selection bias. In this paper, we compare the PSM method with the hand-matching method. This comparison contributes to the growing literature investigating the credibility of the PSM quasi-experimental method for education research (e.g. Aiken, West, Schwalm, Carroll, & Hsuing, 1998; Shadish, Clark, & Steiner, 2008; Wilde, & Hollister, 2007).

Propensity Score Matching

The PSM algorithm calculates the conditional probability of assignment to treatment given a set of observable covariates. The algorithm allows for balancing of the observed covariates, thereby creating a data set similar to what might result from true random assignment. In theory, this should reduce the bias associated with self-selection (Rosenbaum & Rubin 1984).

PSM is well suited to studies in which there are a relatively small number of participants exposed to a given treatment and a relatively large number of participants who did not receive the treatment, from which the final, post-hoc control group can be sampled (Rosenbaum & Rubin, 1985). This documented comparison group condition well describes our data set.

Revisiting a Study of the Effect of Program Enrollment on Student Achievement

In prior research (Tran & Nathan, 2010), we investigated the relationship between enrollment in a high school pre-engineering curriculum—*Project Lead The Way* (PLTW)—and student achievement on state standardized tests. The analysis used multilevel statistical modeling, with students (N = 140) nested within teachers (N = 27). Data obtained from the school district provided information on student achievement in science and mathematics, gender, ethnicity, eligibility for free/reduced-price lunch programs, identification for special education services, and teacher years of experience and academic preparation.

Our sample consisted of 772 students with complete data for the identified factors. The treatment group was identified as those students in the sample who enrolled in one or more of the PLTW high school engineering courses (n = 70). We then handpicked a comparison group (n = 70) from the larger sample that matched the treatment group on three criteria: prior achievement, gender, and free or reduced-price lunch eligibility. This selection technique resulted in two groups of students (N = 140) with comparable course enrollment in science and mathematics, as illustrated in Table 1. The demographics of the study sample are shown in Table 2.

We found that while students gained overall in mathematics and science achievement from 8th to 10th grade, students enrolled in PLTW showed significantly smaller assessment gains in mathematics than those in the hand-matched comparison group that did not enroll; gains of the two groups were comparable in science (Tran & Nathan, 2010).

Research Questions

Hand-matched selection of a comparison group is a difficult and time-consuming method of minimizing selection bias, and its difficulty grows exponentially with the increased dimensionality (number of variables) of the data set. The method is also subjective and error-prone. In addition to addressing issues related to efficiency and reliability, the purpose of this study was to determine if PSM offers a more effective way to select a comparison group. The following research questions guided our comparative analysis of PSM and hand-matching methods:

- Does utilizing a PSM approach to creating a post-hoc comparison group afford the ability to draw inferences from observational data that are at least as robust as those obtained from a hand-matched approach?

- Does use of PSM yield practical benefits?

Method

Sample

Our sample of students was drawn from a Midwestern city with an urban, midsized population (exceeding one-half million). In the 2007–2008 academic year, the district enrolled

more than 87,000 K–12 students, of whom approximately 57% were African American; 22%, Hispanic; 12%, White; 4%, Asian, and 4%, "other." Approximately 72% of the students in the district were eligible for free or reduced-price lunch. Students identified for special education services made up 18% of the student body, greater than the national average (12%), while English language learners made up 8%.

Data Sources

Both mathematics and science assessments were administered to students in November 2005 (8^{th} grade) and again in November 2007 (10^{th} grade). These standardized tests are designed to measure the state academic standards in mathematics and science using multiple-choice and short-answer questions. The scale scores and proficiency categories for mathematics and science are explicitly stated.[1]

The district provided data on student characteristics, including prior achievement on statewide standardized tests in mathematics and science, gender, free or reduced-price lunch eligibility, and course enrollment. These variables, along with 8^{th}-grade achievement in mathematics and science, were included as predictors as described in the following section.

Procedures

We derived propensity scores for the group of students with complete data ($N = 772$), employing the estimation and selection algorithms proposed by Dehejia and Wahba (2002) and the Stata-compliant psmatch2 software package (Leuven & Sianesi, 2003). In the previous study, we analyzed only classrooms with at least five students in them for multilevel analysis, giving us 70 students in the hand-matched treatment group and 70 in the hand-matched comparison group. In contrast, the current study, using single-level analysis, permitted us to lift this restriction, thus giving us a greater number of students in the PSM treatment group. This allowed us to create a new PSM treatment group ($n = 81$; see Table 3) and PSM comparison group ($n = 291$) and determine the frequency of *common support*, an indicator of the degree of distributional overlap. Thus, the PSM treatment group included 11 more cases than the hand-matched treatment group since the single-level analysis removed the constraint for the minimum number of students per school. We then compared the PSM results ($N = 372$) with the results we had obtained using the hand-matching approach ($N = 140$).

The following regression equation was used to estimate the treatment impact on student achievement in science and mathematics:

$$\text{Achievement} = \beta_0 + \beta_1\text{Prior Achievement} + \beta_2\text{Free/Reduced-Price Lunch} + \beta_3\text{Female} + \beta_4\text{Treatment} + \varepsilon$$

Achievement was regressed on the prior achievement score, free or reduced-price lunch status, gender, treatment (PLTW enrollment), and the disturbance term (ε), which captured the influence

[1] The proficiency categories are *advanced*, *proficient*, *basic*, and *minimal performance*. The scale score ranges for mathematics are 350–730 for 8^{th} grade and 410–750 for 10^{th} grade. For science, the scale score ranges are 230–560 for 8^{th} grade and 240–610 for 10^{th} grade.

on student achievement of everything other than the independent variables specified in the equation.

Results

We found that the matched comparison observations that were derived from PSM had a bifurcated distribution that was different from the distribution obtained using the equal weight assumption in the hand-matched scenario, but this differential weighting scheme did not appear to degrade the quality of the model. It is also important to note that a covariate (gender) that was not a significant predictor of student achievement using hand matching emerged as a significant predictor using PSM, suggesting that the PSM algorithm results may be empirically more robust. No differences were found between these two approaches in other predictors.

Table 3 provides a summary of these results. The descriptive statistics show that the four groups—hand-matched treatment (n = 70), hand-matched comparison (n = 70), PSM treatment (n = 81), and PSM comparison (n = 291)—were very similar in prior achievement in mathematics and science, free or reduced-price lunch eligibility, and gender. This suggests that the hand-matched and PSM techniques generated comparison groups comparable not only to the treatment groups but also to each other.

Table 4 shows estimates of the impact of the treatment (PLTW) on student achievement in science and mathematics for the hand-matched comparison group (N = 140) versus the PSM comparison group (N = 372). The dependent variables were the state standardized test scores for the 10^{th}-grade science and mathematics exams. The independent variables included student prior achievement, free or reduced-price lunch status, gender, and treatment. Only the estimates for treatment are shown in Table 4.

Column 1 in Table 4 identifies the subject area (mathematics or science). Column 2 gives an estimate of the effect of the treatment on student achievement in the hand-matched model (N = 140), controlling for prior achievement, free or reduced-price lunch status, and gender. Column 3 provides an estimate of the effect of the treatment on student achievement in the PSM model (N = 372), after controlling for student characteristics. Column 4 indicates whether the signs in the Column 2 and 3 estimates differ. Last, Column 5 shows whether the differences between the hand-matched and PSM estimates are statistically significant.

The signs of the estimated treatment effects using hand-matched and PSM techniques are the same for mathematics but different for science. However, as indicated in Column 5, the differences between these two estimates of the treatment effect on student achievement are not statistically significant, suggesting that the two sampling techniques yield comparable substantive results. The standard error was smaller for PSM in each model, likely due to the larger sample size.

Discussion

The present study suggests that PSM yields results comparable to those obtained via hand matching in a quasi-experimental design using a small number of covariates representing participant characteristics. Hand matching is laborious so the efficiency advantages of PSM

should be even greater with larger sample sizes and a greater number of covariates that are highly correlated with the treatment condition. The PSM comparison group was larger and had less variability, providing greater statistical power. PSM also found an additional significant predictor (gender).

While PSM improves efficiency and statistical power, it has limitations. First, a large sample is required for accurate computation. Second, while PSM can reduce observable biases, significant unobservable biases may remain. Thus, PSM may not match unmeasured individual and contextual variables that may account for treatment effects.

Notwithstanding these limitations, quasi-experimental methods such as PSM and hand matching are necessary because education settings impose constraints on available data and on research design that often make a true experiment impractical. This paper illustrates how PSM can be employed to generate comparison groups that can be used to estimate treatment effects. By showing the efficacy of PSM, we hope to move this form of analysis from the exotic to the commonplace.

References

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., & Hsuing, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review, 22*, 207–244.

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

Dehejia, R. H., & Wahba, S. (2002). Propensity score matching methods for non-experimental causal studies. *The Review of Economics and Statistics, 84*, 151–161.

Leuven, E., & Sianesi, B. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing [Computer software]. Retrieved from http://ideas.repec.org/c/boc/bocode/s432001.html

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., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33–38.

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Tran, N., & Nathan, M. J. (2010). An investigation of the relationship between pre-college engineering studies and student achievement in science and mathematics. *Journal of Engineering Education, 99*(2), 1–15.

U. S. Department of Education, Institute of Education Sciences. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Retrieved from http://www2.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf

Wilde, E. T., & Hollister, R. (2007). How close is close enough? Testing nonexperimental estimates of impact against experimental estimates of impact with education test scores as outcomes. *Journal of Policy Analysis and Management, 26*, 455–477.

Table 1
Course Enrollment of Students in the Original Study Sample (N = 140)

Course	Treatment group (PLTW)[a]	Comparison group (non-PLTW)
Remedial math	0 (0%)	0 (0%)
Core math	68 (97%)	70 (100%)
Advanced math	0 (0%)	0 (0%)
Missing	2 (3%)	0 (0%)
General science	0 (0%)	2 (3%)
Core science	63 (90%)	64 (91%)
Advanced science	0 (0%)	0 (0%)
Missing	7 (10%)	4 (6%)

[a]PLTW = Project Lead The Way.

Table 2
Characteristics of Students in the Original Study Sample (N = 140)

School	Af. Am.[a]	Asian	Hispanic	White	Other	Female	Spec. ed.	FRPL[b]	PLTW[c]
Benjamin									
%	65.3	4.9	26.5	8.2	0.0	42.9	14.3	83.7	73.5
Raw	32	0	13	4	0	2	7	41	36
Hilldale									
%	22.2	25.9	11.1	40.7	0.0	44.4	7.4	48.1	22.2
Raw	6	7	3	11	0	12	2	13	6
Richmond									
%	8.8	1.6	12.5	18.8	0.0	37.5	12.5	68.8	0.0
Raw	11	0	2	3	0	6	2	11	0
Sinclaire									
%	32.2	19.4	41.9	0.0	6.5	45.2	9.7	90.3	48.4
Raw	10	6	13	0	2	14	3	28	15
Western									
%	94.1	5.9	0.0	0.0	0.0	29.4	29.4	88.2	76.5
Raw	16	1	0	0	0	5	5	15	13

Note. School names are pseudonyms. Numbers are rounded to 3 significant digits.
[a]Af. Am. = African American. [b]FRPL = free and reduced-price lunch eligibility. [c]PLTW = Project Lead The Way course enrollment.

Table 3

Descriptive Statistics for Students in the Hand-Matched and PSM Treatment and Comparison Groups, by Core Subject

	Prior achievement Mean (standard deviation)				Free/reduced-price status				Female			
Subject	HMTG (*n* = 70)	HMCG (*n* = 70)	PSMTG (*n* = 81)	PSMCG (*n* = 291)	HMTG (*n* = 70)	HMCG (*n* = 70)	PSMTG (*n* = 81)	PSMCG (*n* = 291)	HMTG (*n* = 70)	HMCG (*n* = 70)	PSMTG (*n* = 81)	PSMCG (*n* = 291)
Mathematics	508.73 (44.15)	509.73 (44.53)	506.78 (46.36)	509.16 (43.28)	0.77 (0.42)	0.77 (0.42)	0.79 (0.41)	0.75 (0.43)	0.41 (0.50)	0.41 (0.50)	0.41 (0.49)	0.45 (0.50)
Science	368.11 (33.36)	371.73 (32.70)	367.79 (32.01)	372.52 (27.95)	0.77 (0.42)	0.77 (0.42)	0.79 (0.41)	0.75 (0.43)	0.41 (0.50)	0.41 (0.50)	0.41 (0.49)	0.45 (0.50)

Note. HMTG = hand-matched treatment group. HMCG = hand-matched comparison group. PSMTG = PSM treatment group. PSMCG = PSM comparison group.

Table 4

Regression Estimates (and Standard Errors) of Treatment Effect: Hand-Matched and PSM Techniques

1	2	3	4	5
	Regression estimate of treatment effect			
Subject	**HM technique (*N* = 140)**	**PSM technique (*N* = 372)**	**Effects opposite in sign?**	**HM/PSM difference significant?**
Mathematics	-8.86† (5.04)	-8.01* (3.33)	No	No
Science	-2.33 (3.99)	1.13 (3.17)	Yes	No

Note. HM = hand-matching. PSM = propensity score matching.
† $p < .10$. * $p < .05$.