Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal.

Volume 11 Number 9, November 2006 ISSN 1531-7714

Consider Propensity Scores to Compare Treatments

Lawrence M. Rudner & Johnette Peyton
Graduate Management Admission Council

The underlying question when comparing treatments is usually whether an individual would do better with treatment X than they would with treatment Y. But there are often practical and theoretical problems in giving people both treatments and comparing the data. This paper presents the use of propensity score matching as a methodology that can be used to compare the effectiveness of different treatments. The method is applied to answer two questions: (1) "Should examinees take a college admissions test near or a few years after graduation?" and (2) "Do accommodated students receive an unfair advantage?" Data from a large admission testing program is used.

The underlying question when comparing treatments is usually whether an individual would do better with treatment X than they would with treatment Y. There are often practical and theoretical problems, however, in giving people both treatments and then comparing data. In program evaluation, for example, it is not practical to subject students to two programs with the same educational goals and have students take essentially the same class twice. Further, exposure to one treatment alters the conditions. The individual being exposed to both treatments is not like the individual exposed to only one treatment.

This paper presents the use of propensity score matching as a methodology that can be used by programs with large amounts of data to compare the effectiveness of different treatments. The method is applied to answer two questions: 1) "Should examinees take a college admissions test near or a few years after graduation?" and 2) "Do accommodated students receive an unfair advantage?" Data from a large admission testing program is used.

Background

Cook and Campbell (1979) describe several widely accepted methodologies for comparing results for different groups. The usual research paradigm consists of the following method:

1. Form treatment and experimental groups, sometimes with a single group serving as its own control.
2. Map treatments to groups.
3. Analyze group differences.
4. Generalize the findings based on groups to tendencies among future individuals.

Defining the groups is a critical first step. This paper provides an example where the seemingly

obvious approach to group formation does not properly address the intended research question. Once the groups are defined, one would want the composition of the groups to be identical. Short of that ideal, statistical adjustments, often in the form of blocking variables or covariate analysis, could be used to adjust for the pre-treatment group differences.

Random assignment of treatment to groups and then comparison of groups is often held as the methodology of choice. In theory, random assignment assures that the groups are identical. Random assignment, however, is not always practical and does not necessarily result in groups that are equivalent in terms of all the important covariates. Rather, with random assignment, the expected values of the covariates over numerous replications are equal. The observed values with one draw are not necessarily equal.

An alternative to random assignment is a matched-pairs design. Each member of the first group is matched with a member of the second group on all the factors the researcher considers to be feasible and relevant. In a well-matched pair, it is as if we are using the same individual twice. When matching is adequate, the variables used for matching that might cause confounding problems are controlled. The approach falls apart when one matches on too few or irrelevant covariates (matching variables), as the match is not necessarily a good one. Matching on many covariates is difficult, especially if one is trying to obtain an exact match when some of the covariates are continuous.

Propensity score matching (Rosenbaum & Rubin, 1985; Rubin, 1997; Joffe & Rosenbaum, 1999) is a refined approach to a matched-pairs design. The covariates are combined to yield a propensity score, and individuals in the treatment group are matched to individuals in the control group based on their propensity score. Using this method, one is weighting the variables by their relative importance and matching based on an optimal composite, rather than by equally weighted individual variables. Further, by matching on many variables, the people receiving the treatments will be quite alike. Rubin (1997) has shown that when one matches on the composite propensity score, the group means and standard deviations on the covariates will also be equivalent.

Example 1: Testing Near or After Graduation

Methodology

Two approaches to answering the question, "Should examinees take a college admissions test near or a few years after graduation?" are examined.

In the first approach, all the examinees taking the test near graduation are compared to all the examinees taking the test after graduation, without regard to possible covariates. Differences in mean admission test scores as well as differences in background characteristics are identified. The implicit question here is, "Do examinees taking a college admissions test near graduation do better than examinees who wait?" This question is not the same as the original question. The examinees are quite different.

The second approach is an application of propensity score matching. Again, differences in mean admission test scores as well as differences in background characteristics are identified. Groups are matched on a variety of covariates using the following procedure:

1. Start with a treatment group taking the test near graduation and a large database of people taking the test later.

2. Draw a random sample from the large database of people taking the test later. This will be control group 1.

3. Run a discriminant function or a logistic regression analysis predicting group membership from a range of covariates (e.g., gender, undergraduate GPA (UGPA), age, years work experience, undergraduate major [dummy coded], desired concentration [dummy coded], and program type [dummy coded]).

4. Compute the probability of being in the treatment group using the discriminant or logistic regression function based on the covariates for everyone in the database. This is the propensity score.

5. Form a new matched pairs control group. For each person in the actual treatment group, compute the propensity score and then find the nearest neighbor in the database, i.e., the person with the closest propensity score. If multiple control group individuals have the closest propensity score, then randomly select from the individuals with the closest scores. Alternately, one could use the caliper approach and find all people in the database whose propensity scores are within a certain, very small, range.

6. One now, theoretically, has samples that are matched, on the margin, on each covariate. Check that assumption by stratifying both the control group and the treatment group into equal-size intervals based on propensity score. The distribution of each covariate within strata should be very close for both groups.

7. The treatment effect is then the difference in the means on the outcome variables (admission test scores) for the two groups.

A nice feature of SPSS is that by selecting the option to output group probabilities, one obtains the propensity scores for all cases, even if only select cases were used to create the equation. We use SPSS to form the propensity scores, sort records based on propensity scores and group membership, move the propensity scores to the first field and the group id to the second field, save the file to disk as a tab separated file, and then use a custom program for form the matched pairs. A SPSS routine to form the matched pairs is available at Raynald's SPSS Tools website, http://pages.infinit.net/rlevesqu/.

Data Source

A database containing 206,852 admission test records for the July 2003 to June 2004 test year formed the initial dataset for the analyses. These records contained test score information as well as a range of background information. "Wait time" was calculated by subtracting undergraduate graduation date from the date of test administration.

Two groups were formed to differentiate 1) examinees taking the test near graduation and planning on enrolling later and 2) examinees taking the test later and planning on enrolling soon after the exam. "Near graduation" was defined as the interval from nine months prior to graduation to two months after graduation. "Planning to enroll later" was defined as planning to wait at least a year after taking the test before enrolling. The contrasting group took the test between 2.1 and 36 months after graduation and indicated that they intended to enroll within 12 months.

Results

Of the 84,470 records in testing year 2004, 2,321 examinees took the admission test near graduation and indicated that they intended to enroll later. Another 39,676 examinees took the test after graduation and planned on enrolling within one year.

Table 1 shows notable differences in the test scores of these examinees. Those examinees taking the test near graduation and waiting to enroll tend to have higher Quantitative, Verbal, and Total test scores. The effect size for taking the test later is about –.23 for the Total test score.

Table 1: Unmatched Groups by Admission Test Scores and UGPA

Score	Near Graduation		After Graduation	
	Mean	**SD**	**Mean**	**SD**
Quantitative	35.02	10.06	32.88	10.16
Verbal	28.08	8.60	26.63	8.68
Total	532.3	114.3	505.8	117.6

As shown in Table 2, however, the groups differ on a number of covariates. Most notably, higher percentages of those taking the test near graduation are business majors, and higher percentages intend to enroll full time. There is also a slight difference in UGPA and citizenship. These differences indicate that perhaps the groups are different in several important ways and, therefore, examinees taking the test later are not a good comparison group for those who take the test near graduation.

By matching on the single propensity score, we were able to form a control group that was quite similar to the treatment group, as shown in Table 3.

Table 2: Characteristics of Unmatched Examinees

Characteristics		**Near Graduation**	**After Graduation**
Gender	Male	55.9%	56.0%
Business Undergraduate Major	Yes	64.7%	54.9%
Intended Enrollment	Full-time	70.9%	58.7%
Citizenship	U.S.	63.0%	61.1%
Undergraduate GPA	Mean	3.34	3.23
Standard Deviation (SD)		.43	.45

Table 3: Characteristics of Matched Examinees

Characteristics		**Near Graduation**	**After Graduation**
Gender	Male	55.9%	56.0%
Business Undergraduate Major	Yes	64.7%	65.1%
Intended Enrollment	Full-time	70.9%	70.6%
Citizenship	U.S.	63.0%	62.6%
Undergraduate GPA	Mean	3.34	3.34
Standard Deviation (SD)		.43	.43

Having matched the groups, we can compare the scores of those who take the admission test near graduation with a similar group of examinees who take the test after graduation. Table 4 presents these differences in achievement test scores. When the After Graduation group demographics are like those of the Near Graduation group, the differences are less pronounced. The effect size is –.16, as opposed to the original –.23.

Table 4: Matched Groups by Admission Test Scores and UGPA

Score	Near Graduation		After Graduation	
	Mean	SD	Mean	SD
Quantitative	35.02	10.06	33.30	10.07
Verbal	28.08	8.60	27.22	8.91
Total	532.3	114.3	513.0	115.9

Example 2: Accommodated Students

Methodology

Two approaches to answering the question, "Do accommodated students receive an unfair advantage?" are examined. The first is an inappropriate examination ignoring the notable differences between accommodated and unaccommodated students on a host of background variables. The second is a propensity score analysis.

Data

The data source for this example was the 1,091,869 individuals who took the GMAT® between July 1, 2001 and March 16, 2006. In that time frame, 4,290 examinees received some form of accommodation.

Though specifics of all these accommodations were not available, data was available from 2005. In that year, approximately 96% of the accommodated GMAT® examinees received additional test time. The other relatively common accommodations included additional break time, special fonts, and special physical accommodations. Approximately 72% of the accommodated examinees received more than one accommodation.

Results

Table 5 shows notable differences in the test scores of these examinees. Those examinees who received an accommodation scored higher Verbal and Total test scores. These differences are both statistically and practically significant.

Table 5: Unmatched Groups by Admission Test Scores and UGPA

Score	Not Accommodated		Accommodated	
	Mean	SD	Mean	SD
Quantitative	35.18	10.34	34.77	9.57
Verbal	27.25	8.83	30.16	8.42
Total	527.2	114.0	546.0	114.1

However, the groups differ on a number of covariates. Table 6 shows a comparison based on a random sample of 15,000 unaccommodated and 2,305 accommodated examinees with complete data. Much higher percentages of accommodated examinees plan to enroll as full-time students, are white, are male, and are United States citizens compared to unaccommodated test takers. Accommodated examinees also tend to be slightly younger and tend to take the GMAT® exam earlier. When evaluated using t-tests at $p < .05$, there are significant differences between the unaccommodated and the accommodated examinees on all of the means and proportions in Table 6, with the exceptions of the percentage of business undergraduates and undergraduate grade point averages.

Table 6: Characteristics of Unmatched Accommodated and Unaccommodated GMAT® Examinees

Characteristics	**Not Accommodated**		**Accommodated**		**Effect Size**
	% Yes	**% No**	**% Yes**	**% No**	
Intend to Enroll Full-time	60.6%	48.9%	75.3%	43.1%	0.31
Plan to Pursue MBA	79.3%	40.5%	82.7%	37.8%	0.08
White	39.9%	49.0%	64.4%	47.9%	0.50
Male	61.0%	48.8%	72.2%	44.8%	0.23
Business Undergraduate Major	44.4%	49.7%	42.0%	49.3%	–0.05
U.S. Citizen	58.1%	49.3%	86.7%	34.0%	0.60
Characteristics	**Mean**	**SD**	**Mean**	**SD**	**Effect Size**
Age	28.19	6.32	27.29	5.07	–0.15
UGPA	3.20	0.50	3.19	0.46	–0.03
Days to Enrollment	209.2	213.6	241.6	256.2	0.15

Only 2,305 of the 4,290 accommodated examinees had complete data on all of the covariates. In order to determine whether listwise deletion would bias the sample, the percentages and means for the 2,305 examinees were compared against the means for all 4,290 accommodated students. T-tests found no significant differences at $p < .05$. All the means and percentages were extremely close.

Discriminant Function Analysis was used to compute propensity scores as a function of the above nine variables using the sample of 15,000 unaccommodated and 2,305 accommodated examinees. The discriminant function was significant ($r=.28$; Wilks' $\lambda=.922$, $df=7$, $p < .05$). The propensity score was then computed for all examinees. Each of the 2,305 accommodated examinees was matched with a randomly drawn unaccommodated examinee with the same propensity score.

Table 7 reveals that the resultant groups were matched quite well. There are no meaningful nor statistically significant differences between the matched groups of accommodated and unaccomodated examinees on any of the nine variables.

The key question is whether accommodated examinees score higher than unaccommodated examinees after controlling for background differences. As shown in Table 8, the mean scores for the 2,305 accommodated examinees and the matched group of 2,305 unaccommodated examinees are virtually identical. None of the differences in the means are statistically or practically significant.

Table 7: Characteristics of Matched Accommodated and Unaccommodated GMAT® Examinees

Characteristics	Not Accommodated		Accommodated		Effect Size
	% Yes	% No	% Yes	% No	
Intend to Enroll Full-time	77.3%	41.9%	75.3%	43.1%	–0.05
Plan to Pursue MBA	80.6%	39.5%	82.7%	37.8%	0.05
White	63.4%	48.2%	64.4%	47.9%	0.02
Male	72.4%	44.7%	72.1%	44.9%	–0.01
Business Undergraduate Major	41.8%	49.3%	41.9%	49.3%	0.00
U.S. Citizen	85.2%	35.5%	86.7%	34.0%	0.04
Characteristics	**Mean**	**SD**	**Mean**	**SD**	**Effect Size**
Age	27.06	5.52	27.29	5.07	0.04
UGPA	3.20	0.46	3.19	0.45	–0.03
Days to Enrollment	245.8	233.1	241.6	256.2	–0.02

Table 8: GMAT® Scores for Matched Groups of Unaccommodated and Accommodated Examinees

Score	Not Accommodated		Accommodated		Effect Size
	Mean	SD	Mean	SD	
GMAT® Verbal	30.3	8.4	30.4	8.2	0.01
GMAT® Quant	34.5	9.6	34.6	9.5	0.01
GMAT® Total	544.8	112.5	546.1	113.1	0.01

Accommodated examinees differ from unaccommodated examinees on a number of important variables, most notably in the percentages of examinees who plan to enroll as full-time students, who are white, who are male, and who are United States citizens. When these and other background differences are taken into account, the GMAT® scores of accommodated and unaccommodated examinees are virtually identical. In other words, when we select a group of unaccommodated examinees who are similar to the accommodated examinees on select variables, their scores are almost exactly the same as the scores of unaccommodated examinees. Had we not controlled for the select variables and just compared accommodated to unaccommodated examinees, we would have drawn a radically different, and erroneous, conclusion.

Summary

For programs that have large amounts of data, propensity score matching can be a powerful approach to data analysis. The technique permits the researcher to address important questions that are often ill-informed by popular techniques. Unfortunately, there are very few applications in education. A search of the ERIC database in March 2006 found only 11 journal articles and eight additional papers referencing propensity score analysis. Only a few of these were applications.

Leow, Marcus, Zanutto, and Boruch (2004) used propensity score analysis to address the difficult question of whether taking advanced courses improves scores on basic achievement tests. Propensity score analysis helped to control for the many systematic differences between students who

choose to take advanced courses and those who do not.

Delander, Hammarstedt, Mansson, and Nyberg (2005) evaluated a pilot training program for immigrants with weak language skills registered as unemployed at public employment offices by matching pilot program participants with non-participants.

Lopez-Acevedo (2003) evaluated the effectiveness of a professional technical education system by comparing graduates with a matched control on a variety of outcome measures.

In each of these examples it was not practical or feasible to use random assignment, and the available comparison groups, while large, clearly differed on critical covariates. Propensity score analysis provided a method to form the groups.

"*If adjustments for the many observed covariates are sufficient to remove the bias in the estimated treatment effects, then adjustments for the single variable, the propensity score, are also sufficient to remove bias*" (Joffe & Rosenbaum, 1999).

However, unlike random assignment of treatment, propensity score matching does little to balance the unobserved covariates. It is critical that one have a set of covariates that have a sound rationale for inclusion and which control for key anticipated biases.

Thus, propensity score analysis is not relevant in all situations. "However, rather than giving up, or relying on assumptions about the unobserved variables, there is substantial reward in exploring first the information contained in the variables that are observed. In this regard, propensity score methods can offer both a diagnostic on the quality of the comparison group and a means to estimate the treatment impact" (Dehejia & Wahba, 1999).

References

Cook, T. D., & Campbell, D. T. (1979). *Quasi Experimentation: Design and Analytical Issues for Field Settings*. Chicago: Rand McNally.

Dehejia, R. & Wahba, S. (1999). Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*, 1053–1062.

Delander, L., Hammarstedt, M., Mansson, J., & Nyberg, E. (2005). Integration of Immigrants: The Role of Language Proficiency and Experience. *Evaluation Review*, *29*(1), 24–41.

Joffe, M.M., & Rosenbaum, P.R. (1999). Propensity Scores. *American Journal of Epidemiology*, *150*, 327–333.

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*(4), 461–478.

Lopez-Acevedo, G. (2003) A Reassessment of Technical Education in Mexico. *Journal of Career and Technical Education*, *19*(2), 59–81.

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.

Rubin, D.B. (1997). Estimating Causal Effects from Large Data Sets Using Propensity Scores. [Supplement]. *Annals of Internal Medicine*, *127*(8S), 757–763.

.

.

Notes

Earlier versions of this paper was presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 2006 and published at http://gmac.com as Graduate Management Admission Council Research Report RR-06-07.

The views and opinions expressed in this paper are those of the authors and do not necessarily reflect the views and opinions of the Graduate Management Admission Council®.

Citation

Rudner, Lawrence M. & Johnette Peyton (2006). Consider Propensity Scores to Compare Treatments. *Practical Assessment Research & Evaluation*, 11(9). Available online: http://pareonline.net/getvn.asp?v=11&n=9

Authors

Lawrence M. Rudner
Johnette Peyton
Research and Development
Graduate Management Admission Council
1600 Tysons Blvd
McLean, VA 22102