USING PROPENSITY SCORES WITH SMALL SAMPLES

William Holmes

University of Massachusetts Boston

Lenore Olsen

Rhode Island College

An earlier version of this paper was presented at the meeting of the American Evaluation Association, San Antonio, 2010.

USING PROPENSITY SCORES WITH SMALL SAMPLES

ABSTRACT

Propensity scores are increasingly being used in large sample studies to control for pre-group differences. Because these scores are often used to match cases, they can result in sample attrition. In smaller sample studies, such attrition leaves too few cases for meaningful analysis. Alternatives when working with small samples are to use propensity scores as covariates to control for pre-group differences and to use propensities as weights in weighted regression.

The use of propensity scores with small samples is compared with the alternative of using baseline measures to control for pre-group differences. The paper also presents a procedure for empirically testing whether construct integrity holds. We use data from a dosage specific study of substance abusing families receiving clinical services and coordinated case management. Program outcomes are examined, comparing the use of propensity scores with the use of time one measures alone.

Results show that propensities can be used as covariates or as weights in small samples and produce more reliable results than matching procedures. Using time one measures as a control produces nearly as good results.

USING PROPENSITY SCORES WITH SMALL SAMPLES

Evaluations of community agency programs often involve the use of comparison groups that have not been created by random assignment. There are many reasons for this. It may be unethical to deny clients service under programmatic or professional standards, or illegal under contractual requirements of insurers or state agencies. Clients may also be in a program by court order. In these instances, evaluators often find that their comparison groups are not equivalent.

Researchers are increasingly recommending that the selection bias built into quasi-experimental studies such as this be controlled through the use of propensity scores (Dehejia & Wahaba, 2002; Foster, 2003; Barth, Gibbons, & Guo, 2006; Guo & Fraser, 2010).   These scores can help to "even out" pre-group differences. Together with controlling for time one measures, they allow us to better understand the independent effect of services on client outcomes.

Propensity scores are the estimated probabilities that subjects will be classified in an intervention, control, or comparison group using only using only information prior to the intervention. The best predictors for this are those pre-intervention factors that differ significantly between the intervention and the control or comparison groups. A pre-intervention measure of the outcome is also often used if it differs between the groups. These variables represent pre-group differences.

Propensity scores are usually used with large samples by matching cases between groups. Propensity matching with large samples has been shown to reduce selection bias that may be present in evaluation designs (Rubin, 1979). It has been noted that with small samples there may be insufficient power to produce meaningful results (Quigley, 2003). Because the client populations in community agency evaluations are often small, there is a need to further examine the ability of propensity scores to produce usable results when working with small samples. This article examines strategies that may be useful in controlling for pre-group differences with small samples.

Data reported in this article were taken from an evaluation of a federally funded program providing clinical case management services to substance abusing families who have also been indicated for child abuse or neglect. It uses a dosage specific design, in which the comparison groups had less involvement in their services than the treatment group (for other examples of dosage specific designs see Foster 2003; Kim & Crutchfield, 2004; Mullins, Bard, & Ondersma, 2005; Nye, Zucker, & Fitzgerald, 1995; Parker et al, 1999). The treatment group was fully involved in their services. The comparison groups were selected from families participating in the program who were moderately or only slightly involved in their services, as determined by the amount of contact they had with the program. Data reported in this analysis were available for a total of 112 families, a sample size not uncommon for community service programs.

In the discussion that follows, we summarize the procedures used to identify variables used to construct propensity scores – parents' marital status, physical health condition, and a summary measure of substance abuse risk, all of which differed significantly among the dosage groups at pre-test.   The paper describes the use of Discriminant Function Analysis to estimate propensity scores when there are more than two groups being compared and discusses issues of group classification when a dosage model is used to specify the groups. It also addresses questions which have been raised as to whether the adjusted outcome indicators from an Analysis of Variance maintain construct validity in comparison with the unadjusted indicators.

Identifying Imbalance and Pre-test Correlates

Differences between the treatment and comparison groups were identified by one-way analyses of variance, as well as crosstabulations where appropriate, between the study groups and pre-test variables. These included demographic and economic characteristics of the families, as well as pre-test scores on the North Carolina Family Assessment Scale (NCFAS) (Reed-Ashcraft, Kirk, & Fraser, 2001), which assesses the risk for child placement, and the Risk Inventory for Substance Abuse-Affected Families (SARI) (Olsen, Allen, & Azzi-Lessing, 1996. The NCFAS includes measures of the parent's housing and financial situations, family violence, parental health, substance abuse, parent-child bonding, parenting skills, and social supports. The SARI assesses several areas of risk, including patterns of substance use, commitment to recovery, effect of use on child caring, and recovery supports. Both measures have high levels of reliability and validity. The NCFAS and SARI scales are completed by staff at intake and again at case closing.

Those variables having statistically significant differences at the .05 alpha level were subject to additional screening. The published literature disagrees as to whether one should include all prior variables as predictors of propensity scores or just those that are significant according to some criteria (Austin et al., 2007; Rubin, 1979). The authors chose to use only those predictors that remained significant when controlling for other predictors, as determined by a Discriminant function analysis. The rationale was that only the remaining predictors add anything significant to the predicted score. Everything else adds random variation.

A series of Discriminant analyses were run and non-significant predictors were excluded. There were only three predictors remaining: marital status (a dummy variable indicating married/cohabiting versus not), the NCFAS pre-test measure assessing the parent's physical health, and a substance abuse risk summary scale from the SARI pre-test assessment.   Because the SARI items are intercorrelated, they were converted to a summated scale based on a Principal Component Factor Analysis that showed there was one underlying component and the score coefficients were similar among the items. The omega reliability of this scale is .79.

The pretest differences between the treatment and the two comparison groups for these three variables are shown in Table 1. Not surprisingly, those with lower levels risk stemming from their substance use were more likely to be involved in services. Single parents were more likely to be highly engaged in services than those who were married or living with a partner. Parents having fewer health-related concerns were more likely to be among those who were either highly or moderately involved in services.

Estimating Propensities

Propensity scores are most commonly estimated using logistic regression or Discriminant Function Analysis(Dehejia & Wahaba, 2002) . When one is comparing only two groups, treatment and comparison or control, logistic regression is used far more often than discriminant function analysis. The latter is used mainly when there are more than two groups; for example, in a dosage response model in which there are more than two dose levels or in a comparison design in which multiple alternative interventions are being assessed.

As Rubin (1979) has demonstrated, it is desirable that the distributions of the confounding variables be normal (or at least symmetrical) to achieve accurate estimates of propensities. The distributions of the confounders also need to overlap between the two groups compared to have propensity scores that are close enough to allow matching. The distribution of the parent physical health measure and the substance abuse risk measure were examined for the three groups to assure that there was overlap among the distributions. Both measures were unimodal and approximately symmetrical. Having determined this, these two variables were then used to predict the propensity scores for the three groups, along with marital status.

The three measures described above were entered in the Discriminant Function Analysisto estimate the propensity scores (the probability of categorization among the diagnostic groups). The first discriminant function accounted for nearly 99% of the variance in classifying cases in the dosage groups. The standardized canonical discriminant function coefficients were used to provide a significance test of the ability of the variables to correctly classify cases into groups. Marital status and parent physical health were statistically significant at the .05 alpha level. The SARI pre-test scale was significant at the .10 level.

Classification function coefficients were used to estimate the probability each case would be classified in each group. Since only the first discriminant function was significant, the probabilities estimated from it were sufficient to classify cases into groups. The other functions were not significant and their probabilities were not used.

When the Analysis of Covariance is used for the outcomes, the pre-test value of the outcome is used as a covariate. Because SARI outcomes at closing are used in this analysis, separate estimates of the propensity scores were made for these measures when using the ANCOVA procedure, deleting the corresponding item at pretest from the estimation whose closing value was being used as an outcome measures. This was done to avoid over controlling variation between the diagnostic groups. In these circumstances separate propensity scores were saved for use in the particular run in which items might have been used twice had this not been done.

Methods of Analysis

Propensity scores were used three ways in the analysis: as covariates, as matching criteria, and as weights in a weighted regression. Effect coefficients were calculated using the results of each of the three procedures, as well as for the procedure when the pre-test value of the outcome measures was used as a control variable. The result of using each of these procedures is then examined to see how similar or dissimilar they are, using outcomes selected from the NCFAS and SARI measures that illustrate the effect of each of these approaches in a small sample situation.

Propensity matching and propensity strata. Propensity scores are used commonly to construct treatment and comparison groups whose members are matched with similar propensity scores or to create sample strata whose propensity scores are within quintiles of the range of scores. In the former case, this creates treatment and control groups whose pre-group differences have been reduced or eliminated. In the latter case, the treatment and comparison groups are compared within the propensity strata, which statistically controls for much of the pre-group difference. These two approaches are referred to as “matching” strategies, in as much as both require finding cases with similar propensity scores.

Both of these uses of propensity scores need relatively large samples, either to find enough cases in each group having similar propensity scores or to have an adequate number of cases in each propensity strata. When one does not have a large number of cases in a sample, it can become difficult to find enough matches for the analysis to produce reliable results.

The matching procedure used in this analysis was to match cases in the treatment and comparison group by similarity of propensity score. A nearest-neighbor matching procedure was used with the restriction that the propensities matched had to be within .05 units of each other (a caliper of .05). This procedure resulted in a set of 37 cases for analysis. The analysis of outcomes reported here examined the differences between the treatment and comparison groups.

Because the design was a dosage group design, there was a distinct pattern among the propensity scores among the groups. High propensity scores in one group (better than .80) were always associated with propensities for cases from other groups that were somewhat lower (much lower than the caliper of .05 that was used). This meant that in using matching procedures, the matched cases came only from cases where the propensities were below .80. As noted by Shadish, Cook, and Campbell (2002) extremely high propensities may represent outliers in a group for which no corresponding match can be found and which do not overlap propensities between groups. It was necessary to discard those cases.

Propensity covariance. Using the propensity scores as a covariate is an alternative way in which they may be used. This eliminates the loss of cases resulting from “unmatched” propensity scores. However, its use has been criticized for having to work with adjusted group means, rather than unadjusted group means (Dobkin et al, 2002; Fraas et al., 2007). It has been said that working with adjusted means may alter the meaning of the construct that is measured. If the meaning of the construct is not altered, however, there is little objection to using propensity scores in this way.

A test of the equality of the variance-covariance matrices of the NCFAS and SARI outcome measures was done, comparing them with the variance-covariance matrices of the adjusted outcome measures to examine whether the constructs for the outcome measures are likely to have changed as a result of adjustment. The chi-square tests of the equality of the two pairs of matrices was not significant, indicating that the constructs represented by the SARI and NCFAS adjusted measures do not significantly differ from those for the unadjusted measures.

Because the analysis showed no significant differences in the equality of the variance-covariance matrices of the adjusted and unadjusted outcome measures, we can assume construct integrity.

The propensity variable was used in the Analysis of Covariance with the outcome variables; along with the dosage group factor and the pretest value of the outcome measure. The General Linear Model (GLM) program in SPSS was used in the Analysis of Covariance (ANCOVA). The effect of dosage group was estimated after controlling for the propensity score and the pretest value of the outcome.

Propensity weighted regression. Propensity scores were used as weights as a third alternative to matching and ANCOVA. Theoretically, a sample weighted properly using propensity scores can remove pretest differences (Freedman & Berk, 2008). The inverse of the propensity score was rescaled to sum to 1 and used to weight the cases in the regression. Busso, DiNardo, and Mcrary (2009) have demonstrated that rescaling is necessary to achieve more accurate weighted regression results. They also found that the published literature is unclear as to the performance of weights compared to matching.

Effect sizes. Effect sizes were calculated using Hedges g (Hedges, 1981). Following this procedure, the difference between the adjusted post test mean scores for the highly involved and slightly involved groups were divided by the pooled within-group standard deviation of the outcome measure. Effect sizes of .8 or greater are considered to be large (Cohen, 1988).

Imbalance reduction. To test whether the propensity scores generated through these three methods reduced or eliminated the pre-test imbalance between the groups, an Analysis of Covariance was run between the three factors and the diagnostic groups, controlling for the propensity score (see Table II). In most instances, the imbalance between groups was greatly reduced or eliminated when the propensity score was controlled. Weighted regression was the least effective of the three approaches in reducing imbalance among the three dosage levels.

Findings

Adding propensity scores to the analysis of program outcomes, where the only control is for the baseline measure, can significantly alter conclusions regarding program efficacy. In Table III, we see the effects of using propensity scores for each of the three different methods, compared with using only the time one measure as a control. With the ANCOVA approach, using propensity scores as control variables, significance levels were typically similar to those produced by the pretest only control, with the exception of the physical abuse measure where significance levels and program effects were considerably weaker using the ANCOVA procedures. Matching procedures produced mixed results, with some measures resulting in significance levels and program effects that were similar to the pretest only controls, but other measures much weaker in their significance levels and program effects. It is possible that due to the smaller number of cases available for analysis, it was more difficult to identify significant trends in the outcome data. Weighted regression produced more variable results, with some measures becoming more significant than when pretest only controls were used, and others becoming far weaker. In several instances, weighted regression also tended to result in far larger program effects than were supported by the other approaches.

Conclusions

This study shows that even though small samples have larger error variances than large sample studies, usable results can be produced using various methods to control for imbalance in comparison groups. It is an empirical question as to whether a given study has sufficient power, design characteristics, and reliability of indicators to produce results that are stable and consistent across different methods for controlling imbalance. In this analysis, we saw that ANCOVA, using the propensity score as a control variable, and matching produced results that were fairly consistent with the pretest only control procedures. Weighted regression procedures resulted in somewhat more variable results. The fact that small samples sometimes produce variable findings should not deter researchers from empirically checking the results that are produced by using multiple procedures. If the results are similar between procedures, they should not be ignored.

REFERENCES

Austin, P.C., Grootendorst, P., & Anderson, G.M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine, 26*, 734–753. DOI: 10.1002/sim.2580.

Bartlett, M.S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, 3(2),* 77-85.

Barth, R. P., Gibbons, C., & Guo, S. (2006). Substance abuse treatment and the recurrence of maltreatment among caregivers with children living at home: A propensity score analysis. *Journal of Substance Abuse Treatment*, *30,* 93-104.

Busso, M., DiNardo, J., & McCrary, J. (2009a). New evidence on the finite sample properties of propensity score matching and reweighting estimators. IZA Discussion Paper 3998. Bonn, Germany: Institute for the Study of Labor.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for non-experimental causal studies. *The Review of Economics and Statistics*, *84 (1)*, 151-161.

Dobkin, P. L., De Civita, M., Paraherakis, A., & Gill, K. (2002). The role of functional social support in treatment retention and outcomes among outpatient adult substance abusers. *Addiction, 97 (3),* 347-356.

Foster, E. M. (2003). Propensity score matching: An illustrative analysis of dose response. *Medical Care, 41 (10),* 1183-1192.

Fraas, J.W., Newman, I., & Pool, S. (2007). A comparison of propensity score analysis to Analysis of Covariance : A case illustration. Presented at the annual meetings of the American Educational Research Association. Chicago, Illinois. 12pps.

Freedman, D. A., & Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation Review, 32 (4),* 392-409. doi: 10.1177/0193841X08317586.

Guo, S. & Fraser, M.W. (2010). *Propensity Score Analysis*. Los Angeles: Sage.

Guo, S., Barth, R. P., & Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review, 28,* 357-383.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6(20),* 107-128.

Kim, S., & Crutchfield, C. (2004). An evaluation of a substance abuse aftercare program for homeless women with children using a confounding variable-control design. *Journal of Drug Education, 34 (3),* 231-233.

Lamb, R., Preston, L., Schindler, C., Meisch, R., Davis, F., Katz, J., Henningfield, J., & Goldberg, S. (1999). The reinforcing and subjective effects of morphine in post-addicts: A dose-response study. *Pharmacology and Experimental Therapeutics, 259(3),* 1165-1173.

Nye, C. L., Zucker, R. A., & Fitzgerald, H. E. (1995). Early intervention in the path to alcohol problems through conduct problems: Treatment involvement and child behavior change. *Journal of Consulting and Clinical Psychology, 63(5),* 831-840.

Olsen, L.J. & Holmes, W.M. (2010). Project Connect: Project Evaluation, October 2009-September 2010. Providence, Rhode Island: Children's Friend and Service.

Olsen, L. J., Allen, D., Azzi-Lessing, L. (1996). Assessing risk in families affected by substance abuse. *Child Abuse and Neglect*, *20(9),* 847-856.

Parker, B., McFarlane, J. Soeken, K., Silva, C., & Reel, S. (1999). Testing an intervention to prevent further abuse to pregnant women. *Research in Nursing and Health, 22 (1),* 59-66.

Quigley, D.D. (2010). Using multivariate matched sampling that incorporates the propensity score to establish a comparison group. CSE Technical Report No. 596. Los Angeles, California: University of California at Los Angeles, Center for the Study of Evaluation.

Reed-Ashcraft, K., Kirk, R.W., & Fraser, M. W. (2001). The reliability and validity of the North Carolina Family Assessment Scale. *Research on Social Work Practice, 11(4),* 503- 520.

Rubin, D.B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74,* 318-328. Stable URL: http://www.jstor.org/stable/2286330.

Rubin, D.R. (2008). *Matched sampling for causal effects.* New York: Cambridge University Press.

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

Table 1: Imbalance at Pre-test

	Dosage Group			
Variables	Treatment	Somewhat	Slight	Total
Marital Status				
Percent Married/cohabiting	11	40	33	20
Significance*				.04
Effect coefficient#				.54
Substance Abuse Risk (SARI)				
Mean score	.36	.42	.51	.41
Significance				.00
Effect coefficient				1.09
Parent Physical Health (NCFAS)				
Mean score	-.21	-.17	-1.04	-.38
Significance				.00
Effect coefficient				.85

*Significance is based on a crosstabulation chi-square test or a one-way ANOVA F-test at .05 alpha level.

#Effect coefficient is Hedge's g.

Table 2: Imbalance Reduction by Method

		Significance				Effect Coefficient*		
Variables	UA	PC	MS	WR	UA	PC	MS	WR
Marital Status	.045	.545	.234	.123	.542	.031	.294	.546
Substance Abuse Risk T1	.000	.954	.540	.000	1.087	.074	.393	1.156
Parent Physical Health	.001	.039	.485	.009	.848	.488	.259	.746
N	116	116	44	127	116	116	44	127

UA, unadjusted; PC, propensity control; MS, matched samples; WR, weighted regression
*Effect coefficient is Hedge's g.

Table 3: Project Connect: Program Outcomes by Dosage Groups and Estimation Method

	Significance				Eta Squared				Effect Coefficient#			
	PO	PC	MS	WR	PO	PC	MS	WR	PO	PC	MS	WR
Child well-being at case closing (NCFAS)*												
Overall well-being	.015	.014	.103	.032	.064	.065	.129	,062	.946	1.062	.696	1.133
Parenting abilities at case closing (NCFAS)												
Overall abilities	.000	.000	.001	.000	.297	.232	.340	.237	1.720	1.591	.1.605	2.661
Family safety at case closing (NCFAS)												
Overall	.029	.058	.040	.000	.064	.096	.169	.175	.649	1.012	.335	1.951
Physical abuse	.109	.402	.518	.134	.056	.008	.074	.073	.990	.385	.659	1.438
Family interaction at case closing (NCFAS)												
Overall interaction	.000	.000	.000	.000	.269	.232	.401	.177	1.890	1.865	1.968	2.150
Substance abuse risk at case closing (SARI)**												
Commitment to recovery	.014	.019	.733	.948	.037	.034	.018	.001	.413	.418	.359	.127
Patterns of use	.053	.055	.326	.958	.025	.024	.064	.015	.313	.326	.097	.150
N	112	112	37	109	112	112	37	109	112	112	37	109

PO, pretest control only; PC, pretest & propensity controls; MS, matched samples; WR, weighted regression
#Hedge's g. *North Carolina Family Assessment Scale (NCFAS) is scored from (1) serious problem to (6) clear strength
**Substance Abuse Risk Inventory (SARI) is scored from (1) low risk to (5) high risk (N =175)