Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 12, Number 8, May 2007 ISSN 1531-7714

Assessing Program Outcomes When Participation Is Voluntary: Getting More Out of a Static-Group Comparison

Robert F. Szafran
Stephen F. Austin State University

This paper describes a straightforward approach to assessing the effect of an educational program when individual student participation in the program is voluntary, pretests are not feasible, and the statistical expertise of program personnel or assessment audiences is limited. Background characteristics of students believed to influence the outcome of interest are selected. In order to compute a control group outcome which can be compared to the program group outcome, control group member outcomes are weighted based on the proportion of program participants with the same combination of background characteristics. In this way the outcomes of the control group are estimated had that control group the identical background characteristics as the program group.

Experimental designs are particularly good at identifying the effect of educational programs on students because individuals are randomly assigned to experimental and control groups, thereby eliminating, or at least reducing, group differences in background characteristics. Unfortunately, ethical and practical reasons regularly prevent the use of experimental designs in assessing educational programs. In lieu of true experimental designs, Campbell and Stanley (1963) recommend quasi-experimental designs and Rossi, Freeman, and Lipsey (1999) endorse multivariate statistical controls; but most quasi-experimental designs require pretest data which are frequently unavailable and many multivariate statistical techniques require a statistical sophistication often not possessed by program personnel and supervisory decision-makers.

While a long-term solution for such a situation is to improve the statistical expertise of those who run and those who oversee educational programs, this article presents a shorter-term solution. It makes use of a relatively weak but commonly used pre-experimental design termed a static-group comparison but strengthens that design with the inclusion of what in the sampling literature is termed poststratification weighting. Using calculations that can be easily performed with any spreadsheet application, the outcomes of program participants are compared to the probable outcomes of non-participants if the non-participants had an identical distribution of background characteristics as the program participants.

A static-group comparison involves a comparison of two groups of individuals on some outcome (Campbell and Stanley, 1963, p. 12). One group has participated in the program to be assessed, the other has not. Membership in the groups was not based on random assignment and, therefore, the groups can be expected to differ on many background traits. No pretest data for the groups are available. The extended illustration used later in this article describes such a situation. A university desires to assess the impact of its first-year seminar program on retention and grade point average. Participation in the seminar program is voluntary. Pretest data cannot be obtained because neither retention nor GPA is measurable at the start of a student's first year in college.

The next section of this paper provides a background for using a matched and weighted control group. That background comes from the experimental literature on matching subjects and the sampling literature on disproportionate stratified random sampling.

MATCHED AND WEIGHTED CONTROL GROUP

In order to improve estimates of population parameters or the effect of experimental treatments, survey researchers and experimenters have long made use of techniques to control variability in related factors. In the case of experimental research, blocking of subjects has been employed (Vogt, 2005, p. 29); in survey research, stratified sampling has been used (Vogt, 2003, p. 312). Both techniques group subjects with similar background characteristics. Blocking and stratification are most effective when combined with random selection – of subjects into treatment groups (experimental designs) or of population elements into a sample (probability sampling). In many cases, of course, random selection is not possible. In these quasi-experimental and non-probability sampling situations, blocking and stratification are still useful (Heckman & Hotz, 1989).

To be effective, blocked designs must block and stratified samples must stratify on characteristics related to the experimental response being measured and the population characteristic being estimated. The experimental literature is particularly good at describing the advantages and limitations of forming comparison groups by means of matching (Babbie, 2004, pp. 226-227; Haslam & McGarty, 2004; Mark & Reichardt, 2004; Rossi et al., 1999, pp. 313-320). While matching cannot, by itself, control for all covariates, careful selection of criteria for matching can reduce error in estimates of treatment effect. The assessment example appearing later in this paper describes the process by which SAT score, high school graduating class rank, and college orientation attendance were selected as criteria by which to match program participants and non-participants.

The dividing of subjects into blocks or of population elements into strata creates both an opportunity and a problem in the presentation of research results. The opportunity is that both overall and subgroup results can be presented – the subgroups being the distinct blocks or population strata from which subjects were assigned or elements selected. The problem is that a method for aggregating the results from different blocks or separate strata must be selected. For this the sampling literature is particularly good because survey researchers are usually concerned to match the heterogeneity in their samples to the heterogeneity that exists in some target population. In order to achieve this, a weighting scheme must often be employed (Vogt, 2003, p. 342).

In stratified sampling, populations are divided into groups based on one or more characteristics believed to affect the topic of primary research interest. In a study of voter candidate preference, for example, the population of eligible voters might be stratified on the basis of gender, race, and income. The researcher then takes steps to ensure that some elements of the population with each combination of traits (for example, African-American middle income females) are included in the sample. Each of these combinations of traits is known as a stratum. In proportionate stratified sampling, the proportion of the sample coming from each stratum perfectly matches that stratum's share of the total population. When proportionate stratified sampling is achieved, the results from the separate strata can be simply combined to provide an overall result because the heterogeneity of the sample matches the heterogeneity of the population.

In disproportionate stratified sampling, strata that correspond to small percentages of the target distribution are usually oversampled and strata corresponding to large percentages of the target distribution are usually undersampled. This is done so that relatively precise statements about each of the strata can be made while keeping total research expenses as low as possible. Because some strata were undersampled while others were oversampled, the characteristics of a disproportionate stratified sample do not match the target distribution. In fact, most attempts at proportionate stratified sampling end up being disproportionate because response rates vary across strata. Differential response rates produce the same effect as under- and oversampling. In both cases, a common response is to employ poststratification weighting so that the sample results from any single stratum carry as much weight in the calculation of the overall result as that stratum's share in the target distribution (Edwards, Rosenfeld, Booth-Kewley, & Thomas, 1997, pp. 125-129; Henry, 1990, p. 28-29; Kish, 2004, pp. 113-14; Orr, 1999, p. 214).

The technique described in this paper mirrors poststratification weighting and might be described as post-program selection weighting. A population of potential program participants is stratified based on characteristics believed to affect those outcomes the program is intended to influence. Within each stratum, some persons choose to participate in the program, others do not. The distribution of participants across the strata constitutes the target distribution. The outcomes of these program participants can be examined at the subgroup (stratum) level or straightforwardly summed to yield an overall result. The outcomes of the non-participants can also be examined at the subgroup (stratum) level but are weighted to match the target distribution before calculating an overall result.

This approach is referred to here as a matched and weighted control group. The non-participants form a

control group; the original division of the population of potential participants into strata constitutes a matching process; and the non-participant results are weighted so that the heterogeneity of the non-participants corresponds to the heterogeneity of the target distribution, that is, the program participants. In this way, program administrators can compare the outcomes of participating individuals to a hypothetical group of non-participating individuals with identical background characteristics. What makes this hypothetical comparison real is that the outcomes for this hypothetical comparison group are based on the actual outcomes of the part of the population which chose not to take part in the program.

PROCEDURE

The creation and use of a matched and weighted control group can be succinctly described. As with most succinct descriptions, however, the procedure becomes clearer when illustrated. The extended example that follows the description of the procedure will hopefully serve that purpose.

1. Identify a small number of important background characteristics believed to influence student outcomes on which program and non-program students differ.
2. Collapse each background characteristic into a small number of categories.
3. Divide the class into strata based on every possible combination of background characteristics.
4. Further divide each of these strata into two subgroups: students who participated in the program and students who did not.
5. For a particular outcome of interest, calculate the overall result for students who participated in the program. This is a simple average or percent.
6. For that same outcome of interest, calculate what would be the overall result for students who did not participate in the program if the number of non-program students in each stratum were identical to the number of program students in that stratum. This is a weighted average or weighted percent. It uses the results of the non-program students but weights them based on the number of program students in each stratum.

AN ILLUSTRATION EXTENDED OVER 11 YEARS

Institutional History

The effects of a first-year seminar program at a regional university in Texas have been assessed using a matched and weighted control group every year since 1994, the year the seminar program was instituted. Enrollment is voluntary in the seminar course which meets twice weekly, carries a single credit, and is graded pass/fail. All students attending summer orientations for new first-year students are encouraged to enroll in the seminar. The academic and social benefits of enrolling in the seminar are emphasized. In the years since the seminar was established, entering first-year classes at the university have ranged in size from 1,754 to 2,380 students. During its first year the seminar enrolled just 13% of the first-year class; but for the last several years it has enrolled about 60% of the class.

Program and university administers have been interested in many of the outcomes of the seminar but none as much as the seminar's effect on 12-month retention and first-year grade point average. The following section describes how the six steps in implementing the matched and weighted control group were done.

Step-by-Step Illustration

***1. Identify a small number of important background characteristic**s.* Program staff began by looking for background characteristics that are related to retention and GPA and on which seminar and non-seminar students differ. The three background characteristics chosen were high school graduating class rank, SAT score, and which, if any, summer new student orientation the student attended. (It is important not to confuse the background characteristics the program staff chose with the method of assessment being described in this article. Other schools assessing other programs might choose to control very different background characteristics.)

All three background characteristics are in the university's database so data collection is simplified. Information on high school graduating class rank is available for all but the few students who were home-schooled or graduated from high schools which do not report class ranking. The university requires entering first-year students to submit SAT or ACT scores. Most submit SAT scores. For students only submitting ACT scores, they were converted to their SAT equivalent (Habley, 1995) for this assessment. All entering first-year students are encouraged to attend a summer new student orientation before starting classes in the fall. Most students do. During the 11 years considered here, the

university offered as few as 4 separate orientations in some years but usually offered 6.

Previous studies at the university had indicated that high school rank, SAT score, and orientation attended were among the best predictors of retention and GPA. Students with high rank in their high school class, high SAT scores, and attendance at earlier orientations were more likely to stay at the university and earn good GPA's. The effect of high school rank, SAT, and orientation attended were stronger than demographic characteristics such as gender, race, or age. Furthermore, rank, SAT, and orientation were not strongly correlated with one another. This meant the three variables were getting at three substantially different background areas.

The reasons why rank and SAT correlate with retention and GPA are reasonably obvious. Why orientation attended affects retention and GPA is not so apparent. Orientation attended probably serves as a proxy for several things. Students more in-the-know about how college in general and registration in particular work, perhaps because of college-experienced family or friends, come to earlier orientations. Students who attend earlier orientations may also be more motivated about attending college. And students who attend no orientation are certainly at a disadvantage in terms of receiving information and advice necessary for college success.

The selection of control variables needs to be carefully considered. Specification error in the form of failure to include background characteristics which distinguish seminar and non-seminar students will result in an incorrect assessment of the program (internal invalidity) if those background differences also impact the outcomes being assessed. The level of initial motivation has always been of particular concern for the assessors of this first-year seminar program. While controlling for the effect of orientation attended may approximate the effect of motivation, it is certainly not a perfect solution. The greater the number of control variables taken into account, the greater the internal validity of the assessment but the more difficult the assessment procedure is to do and to explain to interested parties.

The selection of control variables needs to be done with an eye toward the availability of data. Students with missing data on any of the variables cannot be matched and, therefore, drop out of the analysis. In the case of this university, typically about 4% of entering first-year students have no high school rank and about 1% have no SAT or ACT scores. While the loss of any students from the assessment is regrettable, this was judged an acceptable level of missing cases to proceed with the analysis.

2. Collapse each background characteristic into a small number of categories. For assessing this program at this university, high school rank is coded as top quarter, second quarter, or bottom half. SAT scores are categorized as high (1060 or more), medium (950 to 1050), or low (940 or less). Summer orientation attended is classified as early (attended orientation in first half of summer), late (attended orientation in second half of summer), or none (attended no summer orientation).

Dividing continuous variables into discrete categories is always a judgment call. The smaller the number of categories, the easier the later mathematical computations will be but the less precise the matching becomes. The program staff chose the break points they used sometimes for practical reasons (for example, the categories divide the first-year class into approximately equal size groups) and sometimes for theoretical reasons (for example, students who do not attend orientation represent far less than a third of the first-year class but the failure to attend orientation is known to have a substantial effect on retention and GPA).

3. Divide the class into strata. Three background characteristics each collapsed to just three categories create 27 possible combinations (3 x 3 x 3) of background traits. These 27 combinations form the strata into which the class members are divided. Students in the same stratum have approximately the same high school rank, SAT score, and summer orientation history. The rows in Table 1 describe the 27 strata in this illustration.

4. Further divide each of these strata into two subgroups based on program participation. The students in each of these 27 strata are then divided into two subgroups: those who participated in the seminar and those who did not. In Table 1 columns b through d will include information on the seminar participants and columns e through g will include information on the seminar non-participants.

5. For a particular outcome of interest, calculate the overall result for students who participated in the program. One of the outcomes of interest for this program is 12-month retention. Table 2 shows the calculations for this outcome using the university's fall 2004 entering cohort of new first-year students.

The retention rate expressed as a percent for seminar students is simply

$$\frac{\text{\# of seminar students who returned}}{\text{original \# of seminar students}} \; x \; 100$$

Table 1 Outline for Calculating Matched and Weighted Comparisons

strata			seminar student subgroups			non-seminar student subgroups		
(col. a)			(col. b)	(col. c)	(col. d)	(col. e)	(col. f)	(col. g)
high school rank	SAT score	orientation attended	# in subgroup	subgroup average outcome	b x c	# in subgroup	subgroup average outcome	b x f
top 1/4	1060-1600	early	_____	_____	_____	_____	_____	_____
top 1/4	1060-1600	late	_____	_____	_____	_____	_____	_____
top 1/4	1060-1600	none	_____	_____	_____	_____	_____	_____
top 1/4	950-1050	early	_____	_____	_____	_____	_____	_____
top 1/4	950-1050	late	_____	_____	_____	_____	_____	_____
top 1/4	950-1050	none	_____	_____	_____	_____	_____	_____
top 1/4	400-940	early	_____	_____	_____	_____	_____	_____
top 1/4	400-940	late	_____	_____	_____	_____	_____	_____
top 1/4	400-940	none	_____	_____	_____	_____	_____	_____
second 1/4	1060-1600	early	_____	_____	_____	_____	_____	_____
second 1/4	1060-1600	late	_____	_____	_____	_____	_____	_____
second 1/4	1060-1600	none	_____	_____	_____	_____	_____	_____
second 1/4	950-1050	early	_____	_____	_____	_____	_____	_____
second 1/4	950-1050	late	_____	_____	_____	_____	_____	_____
second 1/4	950-1050	none	_____	_____	_____	_____	_____	_____
second 1/4	400-940	early	_____	_____	_____	_____	_____	_____
second 1/4	400-940	late	_____	_____	_____	_____	_____	_____
second 1/4	400-940	none	_____	_____	_____	_____	_____	_____
bottom 1/2	1060-1600	early	_____	_____	_____	_____	_____	_____
bottom 1/2	1060-1600	late	_____	_____	_____	_____	_____	_____
bottom 1/2	1060-1600	none	_____	_____	_____	_____	_____	_____
bottom 1/2	950-1050	early	_____	_____	_____	_____	_____	_____
bottom 1/2	950-1050	late	_____	_____	_____	_____	_____	_____
bottom 1/2	950-1050	none	_____	_____	_____	_____	_____	_____
bottom 1/2	400-940	early	_____	_____	_____	_____	_____	_____
bottom 1/2	400-940	late	_____	_____	_____	_____	_____	_____
bottom 1/2	400-940	none	_____	_____	_____	_____	_____	_____
			column b sum		column d sum			column g sum
				seminar overall result = (col. d sum) / (col. b sum)			non-seminar overall result = (col. g sum) / (col. b sum)	

Table 2: University Retention Results for Fall 2004 Entering New First-Year Student Cohort

strata			seminar student subgroups			non-seminar student subgroups		
(col. a)			(col. b)	(col. c)	(col. d)	(col. e)	(col. f)	(col. g)
high school rank	SAT score	orientation attended	# in subgroup	subgroup % retained	b x c	# in subgroup	subgroup % retained	b x f
top 1/4	1060-1600	early	171	83	14193	72	82	14022
top 1/4	1060-1600	late	40	83	3320	31	68	2720
top 1/4	1060-1600	none	3	100	300	18	56	168
top 1/4	950-1050	early	76	84	6384	36	61	4636
top 1/4	950-1050	late	26	42	1092	18	61	1586
top 1/4	950-1050	none	1	100	100	9	56	56
top 1/4	400-940	early	93	72	6696	31	71	6603
top 1/4	400-940	late	44	59	2596	10	60	2640
top 1/4	400-940	none	6	17	102	9	56	336
second 1/4	1060-1600	early	43	79	3397	26	73	3139
second 1/4	1060-1600	late	17	76	1292	13	62	1054
second 1/4	1060-1600	none	3	67	201	7	43	129
second 1/4	950-1050	early	76	74	5624	38	66	5016
second 1/4	950-1050	late	35	66	2310	24	58	2030
second 1/4	950-1050	none	2	50	100	12	33	66
second 1/4	400-940	early	132	69	9108	47	57	7524
second 1/4	400-940	late	73	66	4818	34	53	3869
second 1/4	400-940	none	9	78	702	23	35	315
bottom 1/2	1060-1600	early	48	67	3216	20	75	3600
bottom 1/2	1060-1600	late	13	46	598	14	57	741
bottom 1/2	1060-1600	none	2	50	100	5	60	120
bottom 1/2	950-1050	early	36	75	2700	21	52	1872
bottom 1/2	950-1050	late	21	52	1092	10	60	1260
bottom 1/2	950-1050	none	1	0	0	3	67	67
bottom 1/2	400-940	early	33	45	1485	15	67	2211
bottom 1/2	400-940	late	17	59	1003	15	53	901
bottom 1/2	400-940	none	3	67	201	8	38	114
		(column sum)	1024		72730			66795

seminar overall result = (col. d sum) / (col. b sum) = 72730 / 1024 = 71.03

non-seminar overall result = (col. g sum) / (col. b sum) = 66795 / 1024 = 65.23

A more complicated way to get the same result but a way that parallels the calculation to be used for the control group is to compute a weighted percent.

$$\frac{\sum\left(\left(\text{seminar subgroup \% retained}\right)x\left(\text{seminar subgroup size}\right)\right)}{\sum\left(\text{seminar subgroup size}\right)}$$

Using this formula, the spreadsheet multiplies the percent retained for each of the 27 seminar subgroups (column c) by the number of students in the subgroup (column b) and enters the result in column d. The spreadsheet then sums the results in column d and divides that by the total number of students in the seminar subgroups (sum of column b). The seminar students had a 12-month retention rate just slightly over 71%.

6. For that same outcome of interest, calculate what would be the overall result for students who did not participate in the program if the number of non-participants in each stratum were identical to the number of participants students in that stratum. For retention this is done by calculating a weighted percent.

$$\frac{\sum\left(\left(\text{non-seminar subgroup \% retained}\right)x\left(\text{seminar subgroup size}\right)\right)}{\sum\left(\text{seminar subgroup size}\right)}$$

The formula looks similar to the previous one but there is an important difference. While the subgroup retention rates are now for the seminar non-participants, the subgroup sizes are still for the seminar participants. Using this formula, the spreadsheet multiplies the percent retained for each of the 27 non-seminar subgroups (column f) by the number of students in the seminar subgroup (column b) and enters the result in column g. The spreadsheet then sums the results in column g and divides that by the total number of students in the seminar subgroups (sum of column b).

If the students in the fall 2004 entering class who did not take the seminar had background characteristics (at least, high school rank, SAT, and orientation attended) similar to the background characteristics of the seminar students, they would have had a retention rate of about 65%. The seminar students had a retention rate approximately six percentage points higher than the matched and weighted control group.

The university uses this technique each year to assess the effect of the first-year seminar on both 12-month retention and cumulative GPA after two semesters. Only students who return for the second semester of their first year are included in the GPA analysis. This reduces the number of students in the strata and subgroups to the extent that attrition reduces the size of the first-year class between the first and second semesters. The calculations are done exactly as they are in assessing the impact on retention except that a weighted average rather than a weighted percent is produced.

Eleven Years of Assessment Results

When reporting the assessment results, program staff report both simple results which compare participants and non-participants without taking initial differences into account and matched results which take initial differences into account using the procedure described above. It has always been relatively easy to explain what matched results mean. Program staff report that these matched results show how the seminar students would compare to a group of non-seminar students who had approximately the same SAT scores, high school rank, and orientation record as the seminar students. For those students, parents, faculty, or administrators who inquire further about the assessment procedure, the strata, subgroups, and weights are explained.

Table 3 shows the seminar's 11 year assessment history. The top half of the table shows the impact of the first-year seminar on retention, the bottom half on GPA. Both the simple comparison and the matched and weighted comparison of seminar and non-seminar students are shown for each year.

The results of the simple comparisons of seminar and non-seminar students in Table 3 show that the seminar students had higher retention and better GPAs in every year. The matched and weighted results also show that the seminar students always had higher retention and in 9 out of 11 years had higher GPAs but the size of the "effect" of the seminar on students is smaller. This is because the seminar tends to draw students with characteristics more favorable to retention and GPA even before the first-year seminar begins. Put differently, the students who take the seminar are more likely to stay at the university and have higher GPAs even if they never took the seminar. The matched results take this initial advantage into account and, as a result, the "effect" of the seminar is adjusted downwards. Even after taking these background differences into effect, however, the effect of the seminar is positive in 20 of 22 comparisons.

Although the program staff certainly wishes the seminar effect in the matched and weighted results were larger, particularly for the effect on GPA, most believe this reduced effect is closer to the true impact of the course.

Table 3: Eleven Years of Assessment Results

Seminar Advantage in Percent Retained after 12-Months

semester students began	simple comparison	matched and weighted comparison	regression coefficient for seminar[1]
Fall 1994	+12 (66% vs. 54%)	+4 (67% vs. 63%)	0.18
Fall 1995	+ 9 (63% vs. 54%)	+4 (64% vs. 60%)	0.10
Fall 1996	+11 (67% vs. 56%)	+8 (67% vs. 59%)	0.23
Fall 1997	+11 (65% vs. 54%)	+8 (66% vs. 58%)	0.30
Fall 1998	+13 (62% vs. 49%)	+7 (62% vs. 55%)	0.28
Fall 1999	+ 7 (58% vs. 51%)	+1 (59% vs. 58%)	0.08
Fall 2000	+13 (64% vs. 51%)	+8 (64% vs. 56%)	0.35
Fall 2001	+ 6 (60% vs. 54%)	+5 (61% vs. 56%)	0.21
Fall 2002	+10 (64% vs. 54%)	+5 (64% vs. 59%)	0.32
Fall 2003	+11 (71% vs. 60%)	+9 (71% vs. 63%)	0.40
Fall 2004	+ 9 (71% vs. 62%)	+6 (71% vs. 65%)	0.33

Seminar Advantage in Cumulative GPA after Two Semesters

semester students began	simple comparison	matched and weighted control group	regression coefficient for seminar[2]
Fall 1994	+.24 (2.46 vs. 2.22)	+.11 (2.46 vs. 2.35)	0.09
Fall 1995	+.23 (2.35 vs. 2.12)	+.07 (2.35 vs. 2.28)	0.08
Fall 1996	+.11 (2.22 vs. 2.11)	+.07 (2.32 vs. 2.25)	0.06
Fall 1997	+.13 (2.34 vs. 2.21)	+.03 (2.35 vs. 2.32)	0.00
Fall 1998	+.16 (2.40 vs. 2.24)	+.06 (2.41 vs. 2.35)	0.05
Fall 1999	+.04 (2.30 vs. 2.26)	+.02 (2.34 vs. 2.32)	0.04
Fall 2000	+.03 (2.40 vs. 2.37)	+.05 (2.40 vs. 2.35)	0.04
Fall 2001	+.04 (2.33 vs. 2.29)	+.05 (2.34 vs. 2.29)	0.04
Fall 2002	+.06 (2.34 vs. 2.28)	−.03 (2.35 vs. 2.38)	0.03
Fall 2003	+.05 (2.46 vs. 2.41)	−.01 (2.46 vs. 2.47)	0.02
Fall 2004	+.05 (2.46 vs. 2.41)	+.01 (2.47 vs. 2.46)	0.01

[1] Unstandardized coefficient from binary logistic regression of returned/not returned the following fall on SAT score, high school percentile rank, orientation attended, and enrolled/not enrolled in first-year seminar.

[2] Unstandardized coefficient from ordinary least squares linear regression of 2nd semester cumulative GPA on SAT score, high school percentile rank, orientation attended, and enrolled/not enrolled in first-year seminar.

TECHNICAL NOTES

It might seem that the retention rate (and the mean GPA) for the seminar group should be the same regardless of whether a simple comparison or a matched and weighted comparison is done, but they sometimes differ slightly. For example, in the first row of Table 3 the fall 1994 seminar participants are reported to have a 66% retention rate when a simple comparison is done but a 67% retention rate when a matched and weighted comparison is done. This is because some of the seminar students included in the calculation of the simple results drop out of the calculation of the matched results because they lack complete data on the background characteristics. If complete data for all students were present, the simple and matched seminar student results would be identical.

Regression Comparisons

The final column in Table 3 presents the regression coefficients for participation in the first-year seminar when a more traditional multivariate statistical regression analysis is done. The coefficients produced by this more traditional statistical technique correspond well to the differences between the seminar and non-seminar groups using the matched and weighted control group technique. In years when the difference between the groups is large, the regression coefficient is large; in years when the group difference is small, the coefficient is small. For the 11 years for which data are available (N=11), the regression coefficients correlate with the group difference produced using the matched and weighted control group technique at .80 for retention and .77 for GPA. The group differences resulting from the simple comparisons without a matched and weighted control group also correlate positively with the regression coefficients but are not as strong (.50 for retention and .69 for GPA). These results suggest the matched and weighted control group technique is valid.

Subgroups with Few Students

When the entering first-year class is divided into strata and those are then divided into subgroups, some subgroups may have few and possibly no students in them. That is usually not a problem. If one of the program subgroups has no students, then that entire stratum drops out of the analysis. With no students in the program subgroup, the weight for that stratum in calculating the weighted average or percent becomes zero. Similarly, if one of the non-program subgroups has no students in it, the entire stratum must drop out of the analysis because there are no matching students for the control group.

If one of the program subgroups has only a few students in it, that is a problem because the average outcome for the subgroup will be based on only a few persons; but it is actually a "self-correcting" problem. While the subgroup average can be greatly affected by the performance of just one or two students, the small size of the group means the weight assigned to this stratum will also be small.

The only potentially troublesome situation arises when a non-program subgroup has only a few students but the corresponding program subgroup is large. Unlike the previous situation, this is not a "self-correcting" problem. The non-program subgroup average which is based on relatively few students would receive a large weight because the corresponding program subgroup is large. This situation has rarely occurred in assessing the first-year seminar at the university but a working rule has been adopted to drop the entire stratum from the analysis if there are fewer than 10 students in the non-program subgroup and the program students in the stratum outnumber the non-program students by more than a factor of five.

A Complement to More Sophisticated Statistical Procedures

Several multivariate statistical techniques are available to assess outcomes from quasi-experimental designs: Although originally used to reduce error variance in randomized experiments, analysis of covariance (ANCOVA) has also been used to compare treatment groups outcomes when those groups are not randomly created and are known to differ on initial characteristics (covariates) (Myers, 1979, p.406). Multivariate regression, like analysis of covariance, is capable of handling both continuous and categorical independent variables and, in regression's various permutations, can handle both continuous and categorical, normally distributed and otherwise distributed dependent variables (Allison, 1999). Propensity scores can also be used to adjust for initially dissimilar and self-selected treatment groups by assigning subjects a single propensity to participate in the program score based on examination of numerous covariates (Luellen, Shadish, and Clark, 2005).

The use of matched and weighted control groups described in this article is not presented as being statistically superior to any of these other statistical techniques. When audience level of statistical knowledge is high, sophisticated procedures are sufficient. When audience level of statistical knowledge is not high, however, matched and weighted control groups do have a strictly practical advantage. The technique can be easily understood. That advantage does not make it a substitute for more rigorous statistical analysis. Rather, it

can be a complementary approach to be used for presentational purposes when its results are confirmed to be broadly consistent with more sophisticated analytic results.

CONCLUSION

A static-group comparison with a matched and weighted control group is not as good as a randomized experimental design in isolating the effect of a program. It is better, however, than designs that involve no comparison group or comparison groups that take no account of initial differences. While Rossi et al. (1999, p. 265) prefer the use of statistical controls for comparing non-randomly assigned groups, they note that matched control groups are useful when communicating results to audiences unfamiliar with statistical control procedures.

Being able to easily communicate the manner in which a comparable control group was obtained is an advantage that should not be underestimated. The concept of dividing a heterogeneous class into relatively homogeneous subgroups and comparing the effect of the seminar within those subgroups makes sense even to audiences that have little or no statistical sophistication. Listeners convey a sense of comprehension and confidence in the conclusions that rarely appears when conclusions are supported by more statistically sophisticated analyses. At times, less may be more when simpler techniques yield more useful results.

References

Allison, P. D. (1999). *Multiple regression: a primer*. Thousand Oaks, CA: Pine Forge.

Babbie, E. (2004). *The practice of social research* (10th ed.). Belmont, CA: Wadsworth/Thomson.

Campbell, D. T. & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Edwards, J. E., Rosenfeld, P., Booth-Kewley, S., & Thomas, M. D. (1997). *How to conduct organizational surveys: a step-by-step approach*. Thousand Oaks, CA: Sage.

Haslam, S. A., & McGarty, C. (2004). Experimental design and causality in social psychological research. In C. Sansone, C. C. Morf, & A. T. Panter (Eds.), *The Sage handbook of methods in social psychology*. Thousand Oaks, CA: Sage.

Heckman, J. J. & Hotz, V. J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. *Journal of the American Statistical Association*, *84*, 862-880.

Henry, G. T. (1990). *Practical sampling* (Applied Social Research Methods Series volume 21). Newbury Park, CA: Sage.

Kish, L. (2004). *Statistical design for research*. Hoboken, NJ: John Wiley & Sons.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: an introduction and experimental test. *Evaluation Review*, *29*, 530-558.

Mark, M. M., & Reichardt, C. S. (2004). Quasi-experimental and correlational designs: methods for the real world when random assignment isn't feasible. In C. Sansone, C. C. Morf, & A. T. Panter (Eds.), *The Sage handbook of methods in social psychology*. Thousand Oaks, CA: Sage.

Myers, J. L. (1979). *Fundamentals of experimental design* (3rd ed.). Boston, MA: Allyn & Bacon.

Orr, L. L. (1999). *Social experiments: evaluating public programs with experimental methods*. Thousand Oaks, CA: Sage.

Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). *Evaluation: a systematic approach* (6th ed.). Thousand Oaks, CA: Sage.

Vogt, W. P. (2005). *Dictionary of statistics & methodology: a nontechnical guide for the social sciences*. Thousand Oaks, CA: Pine Forge.

Citation

Szafran, Robert F. (2007). Assessing Program Outcomes When Participation Is Voluntary: Getting More Out of a Static-Group Comparison. *Practical Assessment Research & Evaluation*, 12(8). Available online: http://pareonline.net/getvn.asp?v=12&n=8

Note:

The author expresses his appreciation to Randy Swing and the reviewers for comments on an earlier draft of this paper.

Author

Correspondence concerning this paper should be addressed to

Robert F. Szafran
Department of Sociology
Stephen F. Austin State University
Box 13047, SFA Station
Nacogdoches, TX 75962
phone: 936-468-2009
e-mail: rszafran@sfasu.edu