The Use of Propensity Score Analysis to Address Issues Associated with the Use of Adjusted Means Produced by Analysis of Covariance

John W. Fraas Ashland University
Isadore Newman University of Akron
Scott Pool Ashland University

It is common for researchers in the field of education to engage in research that involves two groups (e.g., control and experimental) and not have the opportunity to randomly assign the participants to the groups. The challenge facing educational researchers is how to analyze the differences between two groups when randomization is not possible and selection bias is an issue. An analytic technique commonly used by educational researchers to address this challenge is analysis of covariance. The use of this technique, however, raises two concerns: (a) The inclusion of the covariates in the analysis of the criterion variable may change the construct represented by the criterion variable, and (b) the analytic technique employed does not match the research question, which is a Type VI error. A technique referred to as propensity score analysis, which is also designed to deal with selection bias in the comparison of non-randomized group means, may address these two concerns. How this technique can be applied by educational researchers is presented.

Various analytic methods have been proposed to control nuisance variables in the analysis of two or more groups (Kirk, 1982). One approach is to randomly assign participants to groups. It is common, however, for researchers in the field of education to engage in research that involves two groups (e.g., control and experimental) and not have the opportunity to randomly assign the participants to the groups. As noted by Halperin and Jorgensen (1994), there are two broad classes of research designs where such randomization is not possible (a) observational studies as identified by Cochran (1983) and (b) quasi-experimental designs as discussed by Campbell and Stanley (1963).

McNeil, Newman, and Kelly (1996) stated that "some statisticians . . . take the position that lack of random assignment disallows a meaningful conclusion" (p.155). They further state, however, that it is their "position . . . that research and decisions must be made in the real world. Random assignment of [subjects to] groups is ideal, but insight can be gained when this is not possible" (p.155).

One analytic approach suggested by Kirk (1982) that can be used to analyze data obtained from a non-randomized research design involved the sub-classification of the participants on key covariates and the inclusion of these sub-classifications in the analysis. It should be noted, however, that the sub-classification procedure can be difficult when multiple key covariates are identified (Halperin & Jorgensen, 1994). As the number of covariates increases, the number of sub-classifications grows exponentially. It is quite possible that some of the sub-classifications will contain none of the study's participants.

A second analytic approach often used by educational researchers when random assignment of subjects is not possible is analysis of covariance (ANCOVA) (Huitema, 1980; McNeil, et al., 1996). In analysis of covariance, the researcher estimates and statistically tests the amount of unique variation in the dependent variable accounted for by the variable or variables representing group membership. One concern with the use of ANCOVA to analyze the difference between group means in non-randomized designs, which is the focus of this article, was discussed by Tracz, Nelson, Newman, and Beltran (2005). Tracz et al. stated that:

> It is important to remember that the outcome or dependent variable in ANCOVA is an adjusted score. . . . After the effects of the covariate have been statistically controlled or removed from the dependent variable . . . , the error variance is all that remains. This residualized or adjusted dependent variable is no longer the same as the original dependent variable. (*p*. 20)

Thus, when the covariates are included in the analysis of the criterion variable, the criterion variable may change as a measure of the construct.

A second concern with the use of ANCOVA deals with the lack of congruency between the research question and the analytic technique, which is referred to as a Type VI error (Newman, Deitchman, Burkholder, & Sanders, 1976; Newman, Fraas, Newman, & Brown, 2002). If the research question deals with student achievement, but the analytic technique analyzes adjusted scores due to the inclusion of covariates, the analytic technique may not produce results that directly address the research question.

One technique that may allow researchers to address these two concerns is propensity score analysis. The next section of this article presents the concept of propensity score analysis and its application to a hypothetical set of data.

Propensity Score Methodology

Rosenbaum (2002) and Rosenbaum and Rubin (1983; 1984) presented an analytic method that used propensity scores to adjust the comparison of non-randomized group means for selection bias due to systematic differences on a set of covariates. Rosenbaum and Rubin (1984) stated:

> There exists a scalar function of covariates, namely the propensity score, that summarizes the information required to balance the distribution of the covariates. Specifically, subclasses formed from the scalar propensity score will balance all . . . covariates. In fact, often five subclasses constructed from the propensity score will suffice to remove over 90% of the bias due to each of the covariates. (*p*. 516)

As noted by Yanovitzky, Zanutto, and Hornik (2005):

> The diagnostics and fitting of the propensity score model are done independent of the outcome and, thus, approximate random assignment of the subjects to treatment . . . . Propensity score methods seek to create comparison groups which are similar (or balanced) on all confounders [covariates] but different on their levels of treatment. (*pp*. 210-211)

We believe this characteristic of the propensity score technique allows researchers to address the selection bias concern with respect to the covariates while not changing the construct represented by the criterion variable. In addition, propensity score analysis may allow the researcher to better match the analytic tool and the research question. Thus, a researcher would not be as likely to commit the Type VI error that involves the analysis of adjusted means when the research question of interest deals with unadjusted means.

Steps Used to Conduct Propensity Score Analysis

Propensity score analysis can be understood by reviewing the steps used to conduct such an analysis. The means of conducting propensity score analysis presented in this article is not meant to be an exhaustive discussion of the various ways researchers can implement the technique, but rather the discussion is meant to provide insight into how this analytic technique may allow researchers to address the two concerns previously mentioned regarding the use of ANCOVA.

Yanovitzky et al. (2005) presented six steps researchers may follow to conduct a propensity score analysis.

> *Step 1--Select the covariates*. The researcher must select, a priori, a set of covariates based on theoretical grounds and previous empirical studies. These covariates are used to estimate the propensity scores used to form sub-groups of participants.
>
> *Step 2--Assess the initial imbalance in the covariates*. The researcher gauges the initial imbalance in each of the covariates with respect to the groups. For covariates with interval or ratio level of measurement, an independent-samples t test can be used, while for dichotomous covariates a z test of differences in proportions can be employed. Assessing the initial imbalance in the covariates is useful for two reasons. First, it allows the researcher to determine if the balance is adequate, that is, the degree of balance one would expect in a completely randomized experiment (see Rosenbaum & Rubin, 1984; Zanutto, Lu, & Hornik, 2005). If the balance is adequate, the researcher does not need to employ the propensity score analytic technique and the group means on the criterion variable can be directly analyzed. Second, the assessment of the initial imbalance serves as a benchmark against which the propensity score methodology has increased the balance in the covariates.
>
> *Step 3--Estimate the propensity scores*. If an imbalance exists between the groups with respect to a number of the covariates, the propensity scores are estimated for each of the participants in the study. These propensity scores can be estimated using a variety of methods. Researchers could use discriminant analysis, probit models, or logistic regression models with the dependent variable being the group variable (e.g., control and experimental) and the covariates serving as the independent variables (D'Agostino, 1998; Rosenbaum & Rubin, 1984). McCaffrey, Ridgeway, and Morral (2004) described the use of generalized boosted regression models, which is a multivariate nonparametric regression technique, to estimate the propensity

scores. Yanovitzky et al. (2005) noted that the use of logistic regression models was the most common method used to generate the propensity scores.

Step 4--Stratify the propensity scores. Once the propensity scores are estimated, they are stratified into four or five levels with equal or nearly equal numbers of subjects in the categories. As noted by Cochran (1983), stratifying into more than 4 or 5 groups usually gains very little.

Step 5--Assess the balance on the covariates across the treatment groups. Once the propensity scores are stratified, the researcher needs to verify that the propensity score groups remove any initial bias on the covariates. Yanovitzky et al. (2005) suggested that this verification procedure can be conducted through the use of a two-way analysis of variance (ANOVA), where the two factors are the treatment groups and the propensity score groups and each covariate is used as the criterion variable. Balance is assumed to be achieved when the treatment main effect and the interaction effect are not statistically significant. Yanovitzky et al. noted that:

> If these two conditions are not met, the propensity score should be re-estimated by adding interaction terms and/or non-linear functions (e.g., quadratic or cubic) of imbalanced covariates to the propensity score model. . . . Steps 3 through 5 are repeated until balance is achieved or no further improvement in balance can be made. (*p*. 214)

Step 6--Estimate and statistically test the difference between the treatment means. In this step, the differences between the treatment means on the criterion variable are calculated and statistically tested for (a) each propensity score group and (b) across all propensity score groups. The statistical tests of the difference between the means in each propensity score group can be conducted with the use of t tests..

As noted by Yanovitzky et al. (2005), the overall estimate of the treatment effect is calculated by averaging the differences between means of the treatment groups across all propensity score groups. The overall treatment effect is calculated as follows:

$$\hat{\delta} = \sum_{k=1}^{4} \frac{n_k}{N}\left(\overline{Y}_{ek} - \overline{Y}_{ck}\right) \tag{1}$$

where, $\hat{\delta}$ is the estimated treatment effect; the propensity score groups (1 through 4) are represented by *k*; *N* is the total number of participants; n_k. is the number of participants in the propensity score group *k*; and the means of the criterion variable for the experimental and control groups within a specific propensity score group are $\overline{Y}_{ek}$ and $\overline{Y}_{ck}$, respectively.

The estimated standard error for the estimated treatment effect is calculated as follows:

$$\hat{s}\left(\hat{\delta}\right) = \sqrt{\sum_{k=1}^{4} \frac{n_k^2}{N^2}\left(\frac{s_{ek}^2}{n_{ek}} + \frac{s_{ck}^2}{n_{ck}}\right)} \tag{2}$$

where, n_k is the number of participants in the *k* propensity score group; *N* is the total number of participants; the sample variances of the experimental and control groups are s_{ek}^2 and s_{ck}^2, respectively; and the number of participants in the experimental and control groups are n_{ek} and n_{ck}, respectively.

The *t* value for the estimated treatment effect is calculated by dividing the estimated treatment effect ($\hat{\delta}$) by its standard error ($\hat{S}(\hat{\delta})$).

Table 1. Descriptive Statistics for the Criterion Variable and Covariates

	Treatment Group[a]			
	Control		Experimental	
Variable	Mean	*SD*	Mean	*SD*
Posttest	40.95	7.27	44.91	6.49
Pretest	23.76	8.29	29.51	7.38
OPT	227.82	26.49	231.54	23.57
Ability	116.02	13.89	113.63	11.31

Note: [a]The sample sizes for the control and experimental groups are 123 and 129, respectively.

Table 2. Comparison of Differences between Control and Experimental Groups on Covariates Before and After Propensity Group Formation

	Pre-Propensity Group Formation	Post-Propensity group formation
Variable	*p*	*p*
Pretest	<0.01	0.41
OPT	0.24	0.66
Ability	0.13	0.90

An Illustration of Propensity Score Analysis

To illustrate the application of propensity score analysis, consider a set of hypothetical data collected from a nonrandomized design. For this example the criterion variable (posttest) is a quantitative variable consisting of scores on a math test administered to the students at the completion of the study. Once the criterion variable was identified, the propensity analysis was conducted by completing the six steps previous presented.

Step 1. The following three covariates were identified:

1. The math pretest covariate, which is labeled pretest, consisted of math scores obtained from a test administered prior to the implementation of the instructional methods.
2. The Ohio Math Proficiency Test (OPT) covariate was composed of student scores on the Ohio Math Proficiency Test administered prior to the implementation of the instructional methods.
3. The covariate labeled ability consisted of student scores on a cognitive ability test administered to the students prior to their exposure to the methods of instruction.

In addition to these three covariates, a dichotomized independent variable was formed to identify the instructional method. For this independent variable, which was labeled treatment, the values of zero (control group) and one (experimental group) were assigned indicating the instructional method to which the students were exposed. The mean and standard deviation values of the criterion variable and the three covariates for both the control and experimental groups are listed in Table 1.

Step 2. The initial imbalances between the treatments on the covariates were determined through the use of independent-samples t tests applied to the pretest, OPT, and ability means for the control and experimental groups. The probability values of these three statistical tests are listed in Table 2 under the heading *Pre-Propensity Group Formation*.

Step 3. A logistic regression model was constructed for the purpose of estimating the propensity score for each student. The treatment variable served as the criterion variable for the model and the covariates and their two-way interaction variables were considered as possible predictor variables. The first procedure used in the construction of the model consisted of entering the three covariates (i.e., pretest, OPT, and ability). The next procedure allowed the three two-way interaction variables formed from the three covariates (i.e., pre-by-OPT, pre-by-ability, and OPT-by-ability) to be entered into the logistic regression model in a step-wise fashion. The step-wise procedure used was a forward method of entry with the criterion for entry set at .05 for the probability of the Wald test value of each two-way interaction variable coefficient.

Once the step-wise procedure was completed the final procedure used to construct the logistic regression model involved in constructing the model consisted of a review of the significance levels of the three covariates (i.e., pretest, OPT, and ability). Any covariate with a non-significant coefficient was deleted unless it was used to form a two-way interaction variable that was entered into the model. The results of the analysis of the logistic regression model,

which included the predictor variables of pretest, OPT, ability, pre-by-ability, and OPT-by-ability, are listed in Table 3.

Step 4. The logistic regression model developed in Step 3 was used to estimate a probability for each of the 252 participants in the study. Each probability value represented the probability that the corresponding participant would be a member of the treatment group (i.e., the group assigned a value of one in the treatment variable) based on that participant's covariate values. These 252 probability values, which are referred to as propensity scores, were stratified into four equal groups of 63 participants as follows:

1. A participant with propensity score less than or equal to the first quartile value was placed in Propensity Group 1.
2. A participant with propensity score greater than the first quartile value but less than or equal to the second quartile value was placed in Propensity Group 2.
3. A participant with propensity score greater than the second quartile value but less than or equal to the third quartile value was placed in Propensity Group 3.
4. A participant with propensity score greater than or equal to the first quartile value was placed in Propensity Group 4.

Step 5. Three two-way ANOVA analyses were used to verify that the propensity score groups removed initial bias on the three covariates with the two main effects consisting of (a) the two treatment groups and (b) the four propensity score groups. The probability value of the F test of the treatment main effect for each of the three analyses is listed in Table 2 under the column entitled Post-Propensity Group Formation. Recall that the column in Table 2 entitled Pre-Propensity Group Formation contains the probability values of the statistical tests of the differences between the covariate means for the control and experimental groups for the three covariates. Since the post-propensity group formation probability values are substantially higher than the pre-propensity group probability values, the propensity group formation is considered to have reduced the initial bias between the treatments with respect to the covariates.

As previously stated Yanovitzky et al. (2005) suggested that balance between the treatment groups with respect to the covariates is assumed to be achieved when both the treatment main effect and the interaction effect between the treatment group and propensity group variable are not statistically significant for the analysis of each covariate. As indicated by the p values listed in the column entitled Post-Propensity Group Formation in Table 2, none of the treatment effects was statistically significant for the three covariates. In addition, the p values for the interaction effects between the treatment and propensity group variables in the three two-way ANOVA analyses of the pretest, OPT, and ability variables were .10, .45, and .19, respectively, which indicates that none of the interaction effects was statistically significant.

Step 6. Table 4 contains the posttest mean and standard deviation values for the control and experimental groups for each of the propensity score groups. The corresponding t test values for the four mean differences are also listed. None of the t values corresponding to the four differences between the posttest means of the experimental and control groups reached the critical t value of 1.67, which corresponds to a one-tailed alpha level of .05. Thus, the difference between the control and experimental groups for each of the propensity score groups was not statistically significant.

Since the results of these four t tests resulted in the same conclusion for each propensity group, that is, none of the differences between the control and experimental posttest means was statistically significant, it was appropriate to test the difference between the overall posttest means of the two groups. The overall treatment effect and its standard error values were calculated using Equations 1 and 2, respectively. The treatment effect value was 1.17, which is also equal to the difference between the overall means listed in Table 4, and the standard error value was 0.77. The *t* value for the overall treatment effect (1.52) was calculated by dividing the treatment effect (1.17) by the standard error (0.77). Since this t test value did not reach the one-tailed critical t value of 1.65, which corresponds to

Table 3. Results for the Logistic Regression Model[a]

Variable	Coefficient	Wald	*p*
Pretest	-0.417	2.21	0.14
OPT	0.255	7.72	<0.01[b]
Ability	0.306	5.81	0.02[b]
Pre x Ability	0.005	4.14	0.04[b]
OPT x Ability	-0.002	8.21	<0.01[b]
Constant	-38.375	6.77	<0.01[b]

Note: [a]Δ(-2 Log likelihood) = 68.995, χ^2 = 98.96, df = 5, p <.01, Cox Snell R^2 =.24, Nagelkerke R^2 = .32.
[b]Statistically significant at the two-tailed α = .05.

Table 5. Results of the ANCOVA Analysis of the Posttest Scores Using MLR Model 2[a]

Variable	Coefficient	*t*	*p*
Treatment[b]	1.49	2.63	<0.01[c]
Pretest	0.40	9.00	<0.01[c]
OPT	0.09	6.27	<0.01[c]
Ability	0.07	2.66	<0.01[c]
Constant	1.83	0.64	0.53

Note: [a]R^2 =.692, df_n = 4, df_d = 247, F =139.00, , p<.01, $\overline{R}^2$ = .687
[b]The proportion of unique variation in the posttest variable accounted for by the treatment variable is .009.
[c]Statistically significant at one-tailed α =. 05.

Table 4. Estimated Treatment Effects on Posttest Math Scores Using Propensity Score Groups

Propensity Score Group	Treatment	Group Size	Mean (SD)	*t*
Group 1	Control	51	38.53 (9.15)	0.63[a]
	Experimental	12	40.25 (5.08)	
Group 2	Control	36	41.17 (5.28)	0.17[a]
	Experimental	27	41.44 (7.40)	
Group 3	Control	27	43.56 (4.11)	0.30[a]
	Experimental	36	43.92 (5.10)	
Group 4	Control	9	46.00 (3.97)	1.25[a]
	Experimental	54	48.33 (5.35)	
Overall	Control	123	42.32[b]	1.52[c]
	Experimental	129	43.49[b]	

Note: [a]Not significant at the one-tailed alpha level of .05 (critical *t* value = 1.67 for comparisons within Propensity Score Groups 1 through 4).
[b]The means are the overall estimates averaged over the propensity score groups. The standard error used to calculate the *t* test value for difference between the two overall estimates is 0.77.
[c]Not significant at the one-tailed alpha level of .05 (critical *t* value = 1.65 for overall comparison).

the significance level of .05, the propensity analysis indicated that the overall treatment effect was not statistically significant.

It is important to note that if one or more of the differences between the posttest means of the control and experimental groups were statistically significant, it would not be appropriate to test the overall posttest difference. The researcher would identify the propensity score group or groups for which the differences in the posttest means were statistically significant, and describe the differences of the characteristics of the students in those groups as compared to the characteristics of the students in the propensity score groups for which the differences in the posttest scores were not statistically significantly different.

An ANCOVA Analysis of the Posttest Scores

To better understand how the results of propensity score analysis differ from results generated from an ANCOVA analysis, it is helpful to compare results produced by both analytic techniques. The next section presents an ANCOVA analysis for the data used in the propensity score analysis.

The initial ANCOVA analysis of the posttest criterion variable, which was conducted with the use of a multiple linear regression model (MLR Model 1), included seven independent variables: (a) treatment, (b) pretest, (c) OPT, (d) ability, (e) pre-by-OPT, (f) pre-by-ability, and (g) OPT-by-ability. Since none of the multiple linear coefficients was statistically significant at the .05 level, a second model (MLR Model 2) was constructed and analyzed that did not include the three two-way interaction variables. That is, the amount of variation in the posttest scores accounted for by the three two-way interaction variables in

MLR Model 1 was pooled into the error term in MLR Model 2. The results of MLR Model 2, which included the treatment independent variable and the pretest, OPT, and ability variables as covariates, are listed in Table 5.

The regression coefficient for the treatment variable (1.49) in the MLR Model 2 estimated the difference between the *adjusted* posttest means of the experimental and control groups. The treatment coefficient indicated that the estimated adjusted posttest mean of the experimental group was 1.49 points higher than the estimated adjusted posttest mean of the control group. The *t* test value (2.63) for this coefficient produced a corresponding one-tailed *p* value that was less than .01. Since this one-tailed *p* value was less than the alpha level of .05, the difference between the estimated adjusted posttest means of the experimental and control groups was statistically significant.

To better understand what the *t* test of the treatment variable coefficient produced by the ANCOVA is testing with respect to the posttest criterion variable, it is helpful to note that this test is equivalent to the statistical test (i.e., the *F* test) of the amount of variation in the posttest variable accounted for by the variation in the treatment variable *over and above* the amount of variation accounted for by the covariates. That is, the *t* test of the treatment coefficient is comparable to testing the amount of *unique* variation in the posttest scores accounted for by the treatment variable. It should be noted that statistically testing the unique *proportion* of variation in the posttest scores accounted for by the treatment variable is comparable to testing the unique *amount* of variation accounted for by the treatment variable.

The formula used to convert the *t* test of the treatment coefficient to the proportion of unique variation in the criterion variable accounted for by the treatment variable is as follows:

$$\Delta R^2 = \frac{t^2\left(1-R^2\right)}{df_d} \tag{3}$$

where: ΔR^2 is the proportion of unique variation accounted for by the treatment variable; t^2 is the square of the *t* test value of the treatment coefficient; and df_d is the denominator degrees of freedom value for the multiple linear regression model.

Substituting the *t* test value of the treatment coefficient (2.63), the R^2 value(.692), and the df_d value (247), which were generated by MLR Model 2, into Equation 3 revealed that the proportion of unique variation in the posttest variable accounted for by the treatment variable was .009. This proportion of unique variation accounted for by the treatment variable can be statistical tested with an *F* test. The statistical test of the .009 value produced an *F* test value of 6.92, which is equivalent to the square of the *t* test value for the treatment coefficient. Since the directional *p* value for this *F* test value ($p < .01$) was less than the .05 alpha level, the proportion of *unique* variation in the posttest scores accounted for by the treatment variable was statistically significant, which must be the case when the difference between the adjusted means is statistically significant.

Comparison of Results and Implications

It is interesting to note that this conclusion differs from the results produced by the propensity score analysis. The *t* test results produced by the propensity analysis indicated that no significant differences existed between the mean posttest scores of the experimental and control groups within each propensity score group and across all propensity score groups. The ANCOVA results, however, revealed that the difference between the *adjusted* posttest scores of the experimental and control groups was statistically significant. Why the difference?

One possible reason for the difference in the results of the two analytic methods is the difference in what is being analyzed by the two methods. In the propensity score analysis, the posttest scores of the control and experimental groups were compared within the propensity groups, that is, within groups of students with similar predicted probabilities of being members of the control or experimental groups based on the covariate variables. Thus, the propensity score analysis did not statistically test *adjusted posttest means*, which is to say, the propensity analysis did not test the amount of unique variation in the criterion variable accounted for by the treatment. In contrast, the difference between the *adjusted* posttest means of the control and experimental groups was analyzed in the ANCOVA, which is synonymous with testing the amount of unique variation in the posttest scores (the criterion variable) accounted for by the treatments.

Which method is appropriate? To address this question, two issues should be considered by the researcher. First, if the researcher is concerned that the construct represented by the criterion variable may be substantially altered by the analysis of adjusted means, which is the issue discussed by Tracz et al. (2005), propensity score analysis may provide an appropriate alternative analytic technique to ANCOVA. Tracz et al. note that if ANCOVA is used, researchers should consider establishing reliability and validity estimates for the adjusted scores, which is no small task and, we suggest, unlikely to be done by most researchers. The use of propensity score analysis provides a less time consuming alternative by not analyzing the amount of unique variation in the criterion variable accounted for by the treatments.

Second, if the researcher is interested in testing the difference between the *posttest means* and not the *adjusted posttest means*, propensity score analysis would be the recommended procedure. If, however, the researcher is interested in testing the difference in the *adjusted means*, ANCOVA will provide that analysis. The key point regarding this issue is that the researcher should strive to match the analytic technique to the research question. That is, the researcher should select the research technique that will not lead to a Type VI error.

Summary

The purpose of this article is to suggest the use of propensity score analysis as an appropriate analytical tool for addressing two concerns expressed in the literature. One of these concerns deals with the issue that the construct represented by the criterion variable may change (Tracz et al., 2005) when the analytic tool used by the researcher analyzes *adjusted* means, that is, the amount of unique variation in the criterion variable accounted for by the treatment variable. If ANCOVA is used, Tracz et al. (p. 20) suggest that the "residualized or adjusted dependent variable is no longer the same as the original dependent variable." In such a case, Tracz et al. recommend that the researcher establish the reliability and validity of the residualized or adjusted scores. Since this would be no small task, researchers may find it more practical to utilize propensity analysis to address selection bias issues because it involves the analysis of *means in propensity score groups* rather than the analysis of *adjusted means*, as is the case in ANCOVA.

The other concern deals with the use of an analytic technique that appropriately matches the research question, that is, the researcher avoids committing a Type VI error (Newman, Deitchman, et al. 1976; Newman, Fraas, et al., 2002). Specifically, if a researcher is concerned with selection bias and the research question involves *unadjusted* posttest scores, propensity analysis will produce results that deal with unadjusted posttest scores, while ANCOVA will not.

References

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

Cochran, W. G. (1983). *Planning and analysis of observational studies* (L. E. Moses & F. Mosteller). New York: Wiley.

D'Agostino, R. B. (1918). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17,* 2265-2281.

Halperin, S. & Jorgensen, R. (1994, April). *The use of control in non-randomized design.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA: (ED 369 815).

Huitema, B. E. (1980). *The analysis of covariance and alternatives.* New York: Wiley.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences.* (2nd ed.). Belmont, CA: Brooks/Cole.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods (9)*4, 403-425.

McNeil, K., Newman, I., & Kelly F. J. (1996). *Testing research hypotheses with the general linear model.* Carbondale, IL: Southern Illinois University Press.

Newman, I., Deitchman, R., Burkholder, & J., Sanders, R. (1976). Type VI error: Inconsistency between the statistical procedure and the research question. *Multiple Linear Regression Viewpoints, 6*(4), 1-19.

Newman, I., Fraas, J. W., Newman, C., & Brown, R. (2002). Research practices that produce Type VI errors. *Journal of Research in Education, 12*(1), 138-145.

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

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79,* 516-524.

Tracz, S. M., Nelson, L.L., Newman, I., & Beltran, A. (2005). The misuse of ANCOVA: The academic and political implications of Type VI errors in studies of achievement and socioeconomic status. *Multiple Linear Regression Viewpoints, 31*(1), 19-24.

Yanovitzky, I., Zanutto, E., & Hornik, R. (2005). Estimating causal effects of public health education campaigns using propensity score methodology. *Evaluation and Program Planning (28)*, 209-220.

Zanutto, E. L., Lu, B., & Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national anti-drug media campaign. *Journal of Educational and Behavioral Statistics (30)*1, 59-73.

Send correspondence to: John W. Fraas
Ashland University
Email: jfraas@ashland.edu