October 12, 2006

Improving the Internal Validity of Non-Equivalent Group Research Designs: A Comparison of Propensity Score Analysis and Analysis of Covariance

John Fraas, *Ashland University*
Isadore Newman, *University of Akron*
Joshua G. Bagakas, *Cleveland State University*
David Newman, *Cleveland State University*

Improving the Internal Validity of Non-Equivalent Group Research Designs: A Comparison of Propensity Score Analysis and Analysis of Covariance

John W. Fraas

Ashland University

Isadore Newman

The University of Akron

Joshua G. Bagakas

David Newman

Cleveland State University

A symposium presented at the annual meeting of the

Mid-Western Education Research Association

October 12, 2006

Columbus, OH

Abstract

A challenge facing educational researchers is the need to identify an analytic approach that will (a) analyze the difference between the posttest scores of control and experimental groups in a non-randomized design and (b) appropriately address the research question. A practice frequently 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 analysis of adjusted means does not match the research question, which is a Type VI error. The use of propensity score analysis may address these two concerns. How this technique can be applied by educational researchers and issues related to the application propensity score analysis and ANCOVA are presented.

Improving the Internal Validity of Non-Equivalent Group Research Designs: A Comparison of Propensity Score Analysis and Analysis of Covariance

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. 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 often used by educational researchers when random assignment of subjects is not possible is analysis of covariance (ANCOVA) (Cohen & Cohen, 1983; Hair, Anderson, Tatham, & Black, 1998; Huitema, 1980; Kirk, 1982, McNeil, et al., 1996, Pedahazur, 1973). 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. That is, ANCOVA involves the analysis of the amount of variation in the criterion variable associated with the treatment variable that is independent of the variation in the criterion variable accounted for by the covariates.

A concern with the use of ANCOVA to analyze the difference between group means in non-randomized designs, which is the focus of this paper, 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.

Another 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 paper presents the concept of propensity score analysis and its application to a set of hypothetical data that includes two criterion variables.

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 avoid committing a Type VI error in which the research question dealt with *means* of a given criterion variable and the analytic technique involved the analysis of the *adjusted means* of that criterion variable.

Basis and meaning of propensity score analysis

Propensity score is a conditional probability of a particular observation being assigned to a treatment group given a certain set of observed variables or covariates (Rosenbaum & Rubin, 1984). In a typical quasi-experimental research design, subjects in

treatment groups may differ, not only in the observed values of a set of covariates, but also in the number of covariates. Using a regression-type approach to control for these confounding variables assumes the same number of covariates across treatment groups and therefore may lead to biased results (Yanovitzky, Zanutto, & Hornik, 2005). Thus, using a scalar function of the covariates to produce propensity scores that summarizes each set of the covariates becomes a useful step before any comparisons are made (Rosenbaum & Rubin, 1984). Knowledge of these scores can then be utilized to match subjects based on these scores in order to make more valid comparisons between treatment levels. In other words, propensity score analysis may be viewed as a means of compensating for lack of random assignment in quasi-experimental research.

Estimating the propensity Score

The most common method of estimating the propensity scores in a two-group treatment levels case is using the logistic regression model. The equation takes the form:

$$In\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta X_i$$

Where:

1. P_i is the propensity score.

2. X_i is the asset of covariates used to estimate the propensity score.

Without loss of generality, this conceptualization can be extended to ordinal logistic regression models with more than two treatment levels (Yanovitzky, Zanutto, & Hornik, 2005).

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 amount 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:

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

Where:

1. The estimated treatment effect is $\hat{\delta}$.
2. The propensity score groups (1 through 4) are represented by k.
3. The total number of participants is N.
4. The number of participants in the propensity score group k is n_k.

5. The means of the criterion variable for the experimental and control groups within a specific propensity score group are $\bar{Y}_{ek}$ and $\bar{Y}_{ck}$, respectively.

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

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

Where:

1. The number of participants in the k propensity score group is n^2_k.
2. The total number of participants is N.
3. The sample variances of the experimental and control groups are s^2_{ek} and s^2_{ck}, respectively.
4. 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 ($\hat{s}(\hat{\delta})$).

Propensity Score Analysis: Illustration 1

To illustrate the application of propensity score analysis, consider a set of hypothetical data collected from a nonrandomized design. For this first illustration, the criterion variable (posttest_1) is a quantitative variable consisting of scores on a 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. Three covariates labeled cov_1, cov_2, and cov_3 were identified. 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. In addition to these values, Table 1 contains the posttest_1 mean and standard deviation values for the control and experimental groups.

Insert Table 1 about here

Step 2. The initial imbalances between the treatments on the covariates were determined through the use of independent-samples *t* tests applied to the cov_1, cov_2, and cov_3 means for the control and experimental groups. The results of these statistical tests are listed in Table 2 under the heading *Pre-Propensity Group Formation.*

Insert Table 2 about here

Step 3. The propensity scores were estimated using logistic regression analysis with the treatment variable as the outcome variable and the covariates as predictor variables. The first procedure used in the construction of the model consisted of entering the three covariates. The next procedure allowed the three two-way interaction variables

formed from the three covariates (i.e., cov_1-by-cov_2, cov_1-by-cov_3, and cov_2-by-cov_3) 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. It should be noted that the method used to construct the logistic regression model (i.e., a step-wise method used to enter the two-way interaction terms) is not the only method that can be used.

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., cov_1, cov_2, and cov_3). 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 cov_1, cov_2, cov3, cov_1-by-cov_3, and cov_2-by-cov_3, are listed in Table 3.

Insert Table 3 about here

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

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 (b) and 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 of the statistical test of the difference between the covariate means for the control and experimental groups for each covariate prior to the formation of the propensity score groups. Since the post-formation probability values are substantially higher than the pre-formation probabilities, the propensity score group formation has reduced the initial bias on 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 Table 2 under the column entitled Post-Propensity Formation Group, 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 cov_1, cov_2, and cov_3 variables, which are not listed in Table 2, were .10, .44, and .19, respectively. Thus, none of the interaction effects was statistically significant.

Step 6. Table 4 contains the posttest_1 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 posttest_1 mean differences are also listed. Using an alpha level of .05, none of the *t* values corresponding to the four differences between the posttest_1 means of the experimental and control groups reached the critical value of 1.67, which corresponded to the .05 significance level.

Insert Table 4 about here

Since the results of these four *t* tests resulted in the same conclusion for each propensity group (i.e., none of the differences between the control and experimental posttest_1 means was statistically significant), it was appropriate to test the difference between the overall posttest_1 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 posttest_1 means listed in Table 4. The standard error value was 0.77. The *t* value for the overall treatment effect, which was calculated by dividing the treatment effect (1.17) by the standard error (0.77), was 1.52. The resulting *t* test value of this 1.52 treatment effect value did not reach the critical *t* value of 1.65, which corresponds to the alpha level of .05. Thus, the propensity analysis indicated that the overall treatment effect was not statistically significant for the posttest_1 scores.

ANCOVA: Illustration 1

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 Illustration 1.

The initial ANCOVA analysis of the posttest scores for Illustration 1, which was conducted with the use of a multiple linear regression model (MLR Model 1), included seven independent variables: (a) treatment, (b) cov_1, (c) cov_2, (d) cov_3, (e) treatment-by-cov_1, (f) treatment-by-cov_2, and (g) treatment-by-cov_3. Since none of the multiple linear coefficients for the two-way interaction variables was statistically significant at the .05 level, a second model (MLR Model 2) was constructed and analyzed that did not include these interaction variables. Thus, 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 cov_1, cov_2, and cov_3 as covariates, are listed in Table 5.

Insert Table 5 about here

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 that 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.

Comparison of Results: Illustration 1

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.

Propensity Score Analysis: Illustration 2

The second illustration of propensity score analysis is designed to demonstrate a propensity analysis for a situation in which the testing results of the differences between the control and experimental mean posttest_2 scores for the four propensity groups are not the same. Specifically, in this illustration, the difference between the control and experimental groups for one of the propensity groups is statistically significant, while the differences for the other propensity score groups are not statistically significant.

When conducting the propensity score analysis for the second illustration, the covariates were the same ones as those used in the analysis of the first criterion variable (posttest_1). See Table 1 for the means and standard deviation values for the three covariates and posttest_2 scores. Since the covariate and the treatment variables are the same for both analyses, the results of Steps 1 through 5 obtained from the first propensity score analysis can be applied to the analysis of posttest_2. Thus, the students were assigned to the same propensity groups for the analysis of the posttest_2 scores as they were for the analysis of the posttest_1 scores.

The execution of Step 6 of the propensity score analysis of the posttest_2 scores produced the results listed in Table 6. None of the independent *t* tests of the differences between the posttest_2 means of the control and experimental groups for Propensity Groups 1 through 3 reached the critical *t* value of 1.67. Thus, none of the treatment effects for these three propensity score groups was statistically significant at the one-tailed alpha level of .05. However, the *t* test of the difference between the posttest_2 means of the experimental (71.04) and control (66.22) for Propensity Group 4 ($t = 2.36$) did reach the critical *t* value of 1.67. Thus, the amount by which the mean of the

experimental group exceeded the mean of the control group was statistically significant at the one-tailed alpha level of .05.

Insert Table 6 about here

Since the difference between the posttest_2 means of the control and experimental groups was statistically significant for only one of the propensity groups, it would not be appropriate to test the overall posttest_2 mean difference between the control and experimental groups. That is, the results suggest that the difference between the posttest_2 means of the control and experimental groups exists for certain types of students but not other types of students. Thus, an overall statement regarding the difference between the group means would not be appropriate.

When these types of results are obtained from a propensity analysis it is important for the researcher and evaluators to note that the posttest_2 means differed only for the Propensity Group 4. Logistic regression results could be used to identify future students who would possibly benefit from the treatment effect. That is, the logistic regression coefficients generated by the model used in the study could be applied to a future student's covariate values to determine whether the student would be identified as a member of Propensity Group 4—the group for which the difference between the posttest_2 means of the experimental and control groups was statistically significant.

To illustrate how the estimated logistic regression coefficients could be used to identify students as being the type of students who would benefit from the treatment, consider a student who had the scores of 35, 280, and 130 for the covariates cov_1,

cov_2, and cov_3, respectively. Using the estimated logistics regression coefficients, which are listed in Table 3, and designated covariate values for the student, the predicted log odds ration value (y') would be calculated as follows:

$$y' = -38.375 - 0.417 * 35 + 0.255 * 280 + 0.306 * 130 + 0.005 * (35 * 130) - .002 * (280 * 130)$$

$$y' = -0.85$$

This y' value of -0.85 can be converted to this student's estimated propensity score by using the following:

$$p = e^{y'} / (1 + e^{y'})$$

Where:

1. The letter *e* represents the value for the natural log (2.718).
2. The symbol y' represents the predicted odds log ratio value.

The predicted probability value (propensity score value) for this student is equal to:

$$p = e^{-0.85} / (1 + e^{-0.85})$$

$$p = .30.$$

The probability value of .30 indicates the probability of this student being a member of the treatment group based on the student's covariate values. Since this value does not exceed .71, which was the propensity cut-score for Propensity Group 4, the student would not be considered a member of Propensity Group 4, and thus would not be considered as the type of student who would benefit from the treatment method. The other future students would be identified in the same manner.

ANCOVA: Illustration 2

The initial ANCOVA analysis of the posttest scores for Illustration 1, which was conducted with the use of a multiple linear regression model (MLR Model 3), included

seven independent variables: (a) treatment, (b) cov_1, (c) cov_2, (d) cov_3, (e) treatment-by-cov_1, (f) treatment-by-cov_2, and (g) treatment-by-cov_3. If a coefficient for any of the interaction variables was not statistically significant, it was deleted from the model. The use of this model evaluation criterion resulted in the deletion of the treatment-by-cov_1 and treatment-by-cov_3 variables. The values for the final multiple linear regression model used to conduct the ANCOVA is listed in Table 7. The coefficient for the two-way interaction variable treatment-by-cov_2 ($b = .07$) was statistically significant ($p < .01$). Thus difference between the adjusted means of the control and experimental groups was related to the level of the cov_2 scores.

Insert Table 7 about here

To identify which students would score higher in the experimental group rather than the control group with the ANCOVA, two lines can be derived, plotted, and interpreted from the results of MLR 4. To obtain the regression line for the control group, the value for the treatment variable is set equal to zero. The values for the cov_1 and cov_3 are set equal to their means, which were 23.76 and 116.02, respectively. For the cov_2 variable, two arbitrary values, with one near the bottom and one near the top of the range of values, are selected. The values selected were 170 and 290. With cov_2 set equal to 170, the values were substituted into MLR 4 and the posttest_2 value was calculated as follows:

post_2 = 31.09 -14.768(treatment) + .505(cov_1) +.062(cov_2) +.029(cov_3)
+ .073(treatment-by-cov2)

post_2 = 31.09 -14.768(0) + .505(23.76) +.062(170.00) +.029(116.02)

+ .073(0)(170)

post_2 = 56.99

With cov_2 set equal to 290, the values were substituted into MLR 4, and the posttest_2 value was calculated as follows:

post_2 = 31.09 -14.768(0) + .505(23.76) +.062(290.00) +.029(116.02)

+ .073(0)(290)

post_2 = 64.43

Using these two values generated from the regression line for the control group, a line representing the control group was plotted in Figure 1.

Insert Figure 1 about here

To obtain the regression line for the experimental group, the value for treatment variable is set equal to one. Once again, the values for the cov_1 and cov_3 are set equal to their means, which were 23.76 and 116.02, respectively, and two values used for the cov_2 variable are 170 and 290. With cov_2 set equal to 170, the values were substituted into MLR 4 and the posttest_2 value was calculated as follows:

post_2 = 31.09 -14.768(1) + .505(23.76) +.062(170.00) +.029(116.02)

+ .073(1)(170)

post_2 = 50.59

With cov_2 set equal to 290, the values were substituted into MLR 4, and the posttest_2 value was calculated as follows:

$$post\_2 = 31.09 - 14.768(1) + .505(23.76) + .062(290.00) + .029(116.02) + .073(1)(290)$$

$$post\_2 = 58.19$$

Using these two values generated from the regression line for the experimental group, a line representing the experimental group was plotted in Figure 1.

The interaction effect depicted in Figure 1 reveals that the interaction effect between the treatment variable and the cov_2 variable is disordinal, that is, the lines crossed. The point of intersection of the two lines can be determined by setting the line for the control group equal to the line for the experimental group and solving for cov_2 as follows:

$$31.09 - 14.768(0) + .505(23.76) + .062(cov\_2) + .029(116.02) + .073(0)(cov\_2) =$$

$$31.09 - 14.768(1) + .505(23.76) + .062(cov\_2) + .029(116.02) + .073(1)(cov\_2)$$

$$cov\_2 = 216.98 \approx 217$$

The value of 217 indicates that when the students' cov_2 scores are below 217, the estimated posttest_2 scores are higher in the control group. However, when the students' cov_2 scores are above 217, the estimated posttest_2 scores are higher for the students in the experimental group. Thus, based on the ANCOVA results, educators may consider assigning only students with cov_2 scores above 217 to the treatment method.

Comparison of Results: Illustration 2

For Illustration 2 the propensity score analysis indicated that it would be inappropriate to provide an overall statement regarding the difference between the mean posttest scores of the control and experimental groups due to the fact that the differences between mean posttest scores for the control groups and experimental groups were not

statistically significant for Propensity Groups 1 through 3, while the posttest mean of the experimental group was higher for the experimental group than the control group for Propensity Group 4. Thus, the posttest scores were higher for the experimental group than the control group only for certain types of students, namely students who had covariate values that placed them in Propensity Group 4. These types of students could be identified for future groups of students by estimating their propensity scores through the use of the estimated logistic regression coefficients.

The ANOVA results for Illustration 2 revealed that one could not make a global statement regarding which group would have the highest adjusted posttest mean because of the presence of a statistically significant interaction effect between the treatment and cov_2 variables. By using the diagram containing the interaction effect, one could determine that the adjusted posttest mean of the experimental group was lower than the adjusted posttest mean of the control group for students with cov_2 scores below 217. However, the adjusted posttest mean of the experimental group was higher than the adjusted posttest mean of the control group for students with cov_2 scores above 217.

To further assess the differences between the two analytic techniques for Illustration 2, we compared the specific students identified by both techniques as ones who would benefit from the experimental method. The students identified by the propensity score analysis as benefiting from the experimental method were the ones who had propensity scores that placed them in Propensity Group 4, which required a propensity score greater than .71. The students identified by the ANCOVA as benefiting from the experimental method were the ones who had cov_2 values greater than 217.

The propensity score analysis identified 63 students who should be placed in the experimental method, while the ANCOVA identified 176 such students. A crosstabulation of the students identified by each analytic technique is presented in Table 8. A total of 130 students, which is 51.6% of the total number of students, were classified the same by both methods (i.e., both methods placed 73 students and 57 students in the *Did Not Qualify* and *Qualify* categories, respectively), while a total of 122 students were classified differently by the techniques.

One should be cautious, however, regarding this comparison of the students identified by the two analytic techniques. A possibly more legitimate comparison of the students identified by the two techniques could be made if Johnson-Neyman confidence bands (Johnson & Neyman, 1936) were calculated for the two regression lines generated by the ANCOVA.

Insert Table 8 about here

Choice of Analytic Methods

To address the issue regarding which method should be used, we must first examine a possible reason why the propensity score analyses and the ANCOVAs produced different results? One possible reason for the differences 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.

Possible Limitations with Propensity Score Analysis

Researchers need to be aware of two possible limitations with the use of propensity analysis. First, the size of sample required to conduct propensity score analysis may be quite large. Recall that the propensity groups are formed by ordering the propensity scores estimated by the logistic regression model in which the treatment variable serves as the criterion variable and the covariates are included as the predictor variables, and then dividing the propensity scores into the propensity groups.

When the covariates differ between the control and experimental subjects, the numbers of control and experimental students in a given propensity group, especially the lowest and highest groups, may differ substantially. If the total sample size is small, the number of students in one of the groups (i.e., control group or experimental group) may be small enough to raise concern regarding statistical testing or generalizability. A review of Table 4 reveals that even with a total sample size of 252, the experimental group in Propensity Group 1 contained only 12 students, and the control group in Propensity Group 4 contained only 9 students. If a study used a relatively small sample, some of the subgroups in the lowest and highest propensity groups may be prohibitively small.

Second, the use of propensity analysis in a situation in which the covariate variables are highly confounded with the treatment variable may be problematic. When the variables are highly confounded, the degree of generalizability of the propensity results may be considerably limited. That is, the people who exhibit the combinations of

covariate values that allow the control and experimental groups to be equated may be very limited in number in the population. Thus, the confounding issue is a design issue and not one that can be addressed by analytic means.

Suggestions for Future Research on Propensity Score Analysis

We believe two topics should be pursued as future avenues of investigation research with respect to propensity score analysis and ANCOVA. One such avenue deals with the issue of whether the students who were identified as "benefiting" from the experimental method by propensity score analysis, such as in Illustration 2, are more likely to match those students identified by ANCOVA when the Johnson-Neyman technique is applied to the regression lines. When the Johnson-Neyman technique is applied to the regression lines in the ANCOVA the covariate score beyond which the mean scores of the two groups are *statistically significant* is identified. An examination of the application of the Johnson-Neyman technique to the ANCOVA regression lines would provide additional insight regarding the degree of match between the groups of students identified as "benefiting" from exposure to a given method by the propensity score analysis and ANCOVA.

Another avenue of future investigation deals with the different nature of the interaction effects as investigated by the two analytic methods. To illustrate the differences between the nature of the interaction effect investigated by propensity score analysis and ANCOVA, consider a case in which two two-way interaction effects were statistically significant in the ANCOVA analysis. With ANCOVA the two-way interaction effects would be interpreted individually, while with propensity score analysis

the interaction effects would be interpreted together, that is, in a more *global* fashion due to the nature of the propensity group formation.

One avenue of investigation is to determine if multiple regression models can be designed to conduct ANCOVA that will produce a more global interpretation of the interaction effects. Such models would use a covariate that does not represent one interaction term but all such interaction terms between the covariates and the treatment variable.

To illustrate, assume we have three covariates labeled cov_1, cov_2, and cov_3 along with an independent variable labeled treatment, and the dependent variable is labeled post_3. Also assume we would like our regression models used to conduct the ANCOVA to incorporate any two-way interaction effects between the covariates and the treatment variable.

The first multiple linear regression model would be constructed as follows:

$$post\_3 = a_0 + b_1(treatment) + b_2(cov\_1) + b_3(cov\_2) + b_4(cov\_3) + \varepsilon_1$$

Where:

1. The symbols a_0 and ε_1 represent the constant and error terms, respectively.
2. The symbols b_1 through b_4 represent the partial regression weights for the treatment and covariate variables.

Once the partial regression weights are obtained, a variable, which is labeled com_cov, is computed by summing the products of each covariate and its corresponding partial regression weight. That is, com_cov is calculated as follows:

$$com\_cov = b_2(cov\_1) + b_3(cov\_2) + b_4(cov\_3)$$

Once this new variable is created, a second new variable, which is labeled com_cov-by-treatment, is created by multiplying the com_cov variable by the treatment variable.

After this interaction variable is generated, a second multiple linear regression model is constructed and analyzed. In this model the post_3 scores are regressed onto the treatment, com_cov, and com_cov-by-treatment variables. Thus, this second model is as follows:

$$post\_3 = a_0 + b_1(treatment) + b_2(com\_cov) + b_3(treatment\text{-}by\text{-}com\_cov) + \varepsilon_2$$

A *t* test of the b_3 coefficient would indicate whether the *global* interaction effect was statistically significant. If it was statistically significant, the partial regression coefficients in the second model could be used to generate a graph similar to the one contained in Figure 1. In addition, the Johnson-Neyman confidence bands could also be calculated for this interaction effect.

There is an important difference between the type of interaction effect modeled by this procedure and the one normally modeled by regression models used to conduct ANCOVAs. The type of interaction contained in Figure 1, which was used for Illustration 2, allows a researcher to interpret only one two-way interaction at a time. The graph generated by the technique presented in this section of the paper would allow a researcher to examine the interaction effects globally and place students into groups accordingly. Such an ANCOVA analysis may be more inline with the question being analyzed by propensity score analysis. Future investigation and discussion of this topic would be beneficial to educational researchers and evaluators.

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 concern is that when adjusted means are being analyzed, the construct represented by the criterion variable may change (Tracz et al., 2005). 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 will appropriately match the research question, that is, avoid 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.

It is important for researchers to realize that the results produced by propensity score analysis and ANCOVA may not result in the same conclusions. Thus, researchers must give significant thought to the reasons for applying a given analytic technique.

Reference

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

Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Hair, J. H., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

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

Johnson, P. O. & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. In J. Neyman and E. S. Pearson (Eds.), *Statistical Research Memoirs, 1936, 1,* 57-93.

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.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). Fort Worth, TX: Harcourt Brace College Publishers.

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.

Table 1

Descriptive Statistics for the Criterion Variables and Covariates

Variable	Treatment Group[a]			
	Control		Experimental	
	Mean	*SD*	Mean	*SD*
Posttest_1	40.95	7.27	44.91	6.49
Posttest_2	60.47	7.03	65.65	7.73
Cov_1	23.76	8.29	29.51	7.38
Cov_2	227.82	26.49	231.54	23.57
Cov_3	116.02	13.89	113.63	11.31

[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

Variable	Pre-Propensity group formation	Post-Propensity group formation
	p	*p*
Cov_1	<.01	.41
Cov_2	.24	.66
Cov_3	.13	.90

Table 3

Results for the Logistic Regression Model[a]

Variable	Coefficient	Wald test value	*p*
Cov_1	-0.417	2.21	.14
Cov_2	0.255	7.72	<.01[b]
Cov_3	0.306	5.81	.02[b]
Cov_1-by- Cov_3	0.005	4.14	.04[b]
Cov_2-by- Cov_3	-0.002	8.21	<.01[b]
Constant	-38.375	6.77	<.01[b]

[a]Δ(-2 Log likelihood value) = 68.995, $\chi^2 = 98.96$, $df = 5$, $p < .01$, Cox and Snell $R^2 = .24$, Nagelkerke $R^2 = .32$.

[b]Statistically significant at the two-tailed alpha level of .05.

Table 4

Estimated Treatment Effects on Posttest Math Scores Using Propensity Score Groups

Propensity Score groups	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]	

[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).

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	<.01[c]
Cov_1	0.40	9.00	<.01[c]
Cov_2	0.09	6.27	<.01[c]
Cov_3	0.07	2.66	<.01[c]
Constant	1.83	.64	.53

[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 the one-tailed .05 alpha level.

Table 6

Estimated Treatment Effects on Posttest_2 Scores Using Propensity Score Groups

Propensity Score groups	Treatment	Group size	Mean (SD)	*t*
Group 1	Control	51	57.86 (8.72)	0.40[a]
	Experimental	12	58.92 (5.43)	
Group 2	Control	36	60.72 (4.91)	0.04[a]
	Experimental	27	60.78 (7.34)	
Group 3	Control	27	63.15 (3.96)	0.25[a]
	Experimental	36	63.47 (5.80)	
Group 4	Control	9	66.22 (4.18)	2.36[b]
	Experimental	54	71.04 (5.87)	
Overall	Control	123	61.99[c]	1.97
	Experimental	129	63.55[c]	

[a]Not significant at the one-tailed alpha level of .05.

[b]Significant at the one-tailed alpha level of .05.

[c]The means are the overall estimates averaged over the propensity score groups. The standard error used to generate the *t* test value for difference between the two overall estimates is 0.79.

Table 7

Results of the ANCOVA Analysis of the Posttest_2 Scores (MLR Model 4)[a]

Variable	Coefficient	*t*	*p*
Treatment[b]	-14.77	-2.85	<.01[b]
Cov_1	0.51	10.52	<.01[b]
Cov_2	0.06	3.01	<.01[b]
Cov_3	0.03	0.99	.32
Treatment-by-cov_2	0.07	3.27	<.01[b]
Constant	31.09	8.05	<.01[b]

[a]$R^2 = .70$, $df_n = 5$, $df_d = 246$, $F = 116.41$, $p < .01$, $\overline{R}^2 = .69$

[b]Statistically significant at the .05 level.

Table 8

Crosstabulation of Students Identified as the Type Who Should Receive the Experimental Treatment by the Two Analytic Techniques

	Classification of students with ANCOVA	
Classification of students with propensity score analysis	Did not qualify	Qualify
Did not qualify	73[a]	116
Qualify	6	57[a]

[a]A total of 130 students (73 + 57) received the same classification by both methods (51.6%).

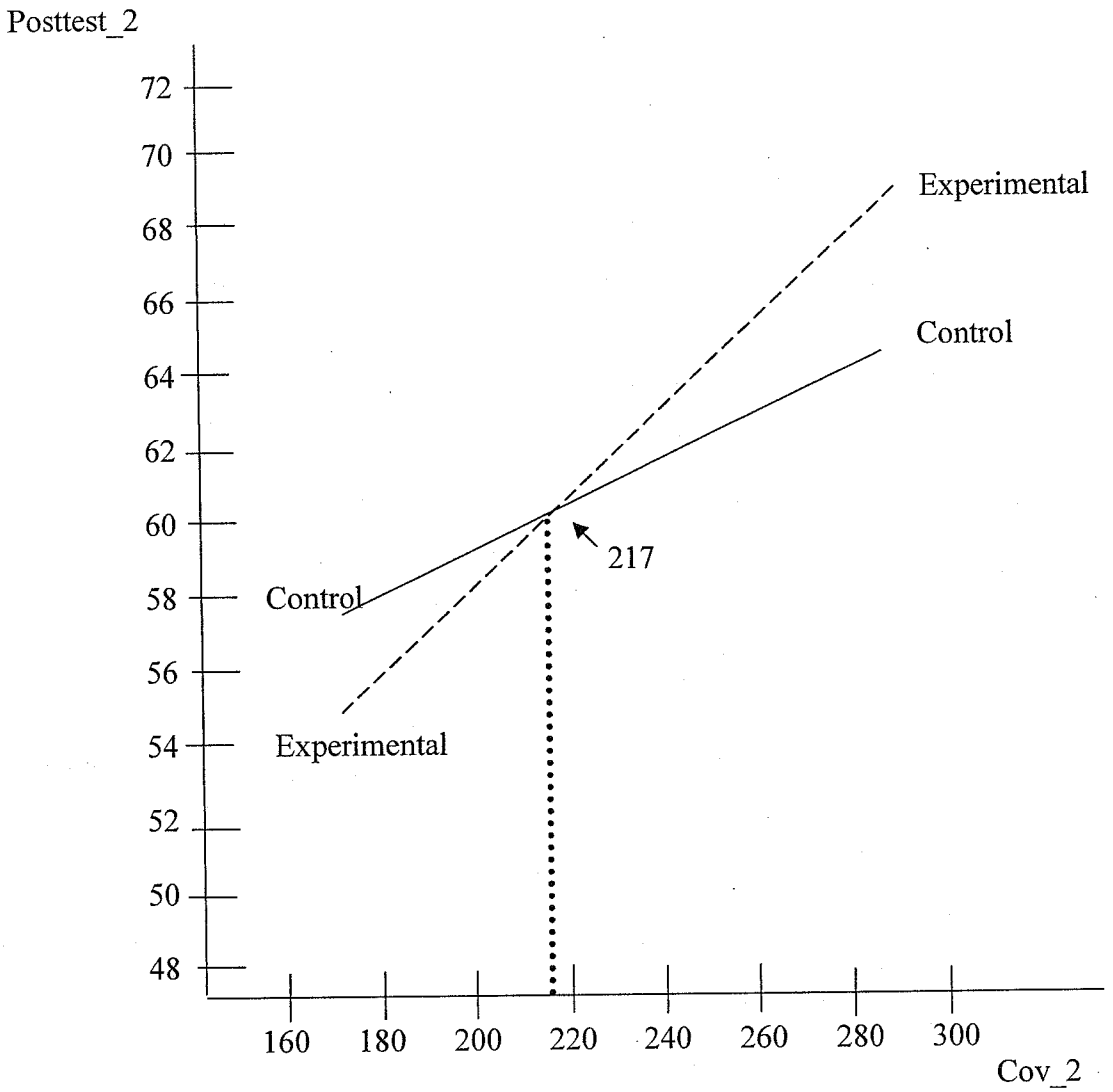

Figure 1. *Interaction Effect for Illustration 1*