Propensity Score Analysis and Analysis of Covariance Results: Do They Address the Same Research Question?

John Fraas, Ashland University
Isadore Newman, University of Akron

Available at: https://works.bepress.com/john_fraas/10/
Propensity Score Analysis and Analysis of Covariance Results:
Do They Address the Same Research Question?

Isadore Newman
The University of Akron
John W. Fraas
Ashland University
David O. Newman
Cleveland State University
Lisa Burg
Susan Shreve
Ashland University

Paper presented at the annual meeting of the Eastern Educational Research Association
Clearwater, FL
February 14-17, 2007
Abstract

Analysis of covariance is an analytic technique frequently used to control for selection bias when random assignment of subjects to the groups is not feasible. A number of concerns have been raised with respect to the results produced by this analytic technique. Recently, educational researchers have used propensity score analysis to deal with the issue of selection bias. This paper compares the results produced by the application of these two analytic techniques to a hypothetical set of data that reflects a differing treatment effect for certain types of students, that is, an interaction effect. Although the results of both techniques were somewhat similar, differences in the results did exist with respect to the numbers of students who "benefited" from the experimental method. The propensity score analysis identified 63 such students, while the ANCOVA identified 171 students. A possible explanation for the lack of agreement with respect to the numbers of students identified by the techniques, which focuses on the difference between the dependent variables analyzed by the techniques, is discussed using the Ballantine Approach and multiple linear regression models.
Propensity Score Analysis and Analysis of Covariance Results:

Do They Address the Same Research Question?

One of the most meaningful questions one can ask in research is: How do the criterion scores differ between various treatments? To answer this question effectively, one needs to control potential alternative hypotheses that can explain variability in the criterion variable. The most effective way to control for alternative hypotheses is to have a research design in which the researcher randomly places subjects in the experimental and control groups, that is, a randomized design (Campbell & Stanley, 1966; Cook & Campbell, 1979, Shadish, Cook, & Campbell, 2002). The use of such a research design allows the researcher to assume that any differences between the experimental and control groups are due to the treatment effect since randomization controls for any differences between the groups at a specified probability level other than treatment.

Unfortunately, in much of the research conducted in the behavioral and social sciences random assignment is not feasible. One of the most frequently used techniques to control for alternative hypotheses is analysis of covariance (ANCOVA) (Cohen & Cohen, 1983; Hair, Anderson, Tatham, & Black, 1998; Huitema, 1980; Kirk, 1982, McNeil, Newman, & Kelly, 1996; Pedahazur, 1982). Subcategories of ANCOVA have been identified as using part correlation, partial correlation, and semipartial correlation techniques.

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 the criterion variable is different from the original criterion variable.

The fact that the ANCOVA produces analyses of adjusted means may open the possibility that the results do not address the research question. Such a lack of congruency between the analytic technique and the research question was referred to as a Type VI error by Newman, Deitchman, Burkholder, and Sanders (1976) and Newman, Fraas, Newman, and 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 does not produce results that directly address the research question.

One technique that may allow researchers to address data produced by non-randomized designs without using adjusted mean scores of the criterion variable is propensity score analysis (Rosenbaum, 2002; Rosenbaum & Rubin, 1983, 1984; Yanovitzky, Zanutto, & Hornik, 2005. As noted by Yanovitzky et al.:

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)

Fraas, Newman, Newman, and Bagakas (2006) discussed and presented an illustration how 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, which was the issue posed by Tracz et al. (2005). That is, 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.

This paper, which expands on the discussion of the comparison of propensity score analysis and ANCOVA results presented by Fraas et al. (2006), presents a comparison of the results produced by propensity score analysis and ANCOVA applied to a hypothetical set of data. The hypothetical data set used in this paper, which is assumed to be generated from a non-randomized design, reflects a differing treatment effect for certain types of students, that is, an interaction effect. To conduct a legitimate comparison of the results produced by the two techniques, regression models were employed in the ANCOVA that allowed a global interaction effect between the covariates and the treatment effect to be estimated and statistically tested. In addition, Johnson-Neyman confidence limits for the global interaction effect were calculated (Fraas & Newman, 1997; Johnson & Neyman, 1936; Pedhazur, 1973).
Application of Propensity Score Analysis and Analysis of Covariance

To compare the results of propensity score analysis and ANCOVA, both analytic techniques were applied to a data set in which the researchers are attempting to determine whether one or more interaction effects exist between a treatment variable, which is a dichotomized variable representing two instructional methods, and three covariates. As previously noted, the hypothetical data used in this paper are assumed to be generated from a non-randomized design. That is, the students were not randomly assigned to the two instructional groups utilized in the study.

The hypothetical data set consists of a criterion variable and three covariates labeled cov_1, cov_2, and cov_3. In addition to the 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.

Insert Table 1 about here

Assume the researchers are interested in addressing the two research questions. First, do interaction effects exist between the covariates and the treatment variable when analyzing the criterion scores, and if so how should the interaction effect be interpreted? Second, if interaction effects do not exist, do the criterion means of the groups differ
controlling for differences between the groups with respect to the covariates? To address these questions, researchers would typically use ANCOVA as the method of analysis. Researchers could also use propensity analysis to address these questions. The following sections of the paper present (a) an application of propensity analysis to a hypothetical data set, (b) an application of ANCOVA to the same hypothetical data set, (c) a comparison of the results generated by the two analytical techniques, and (d) a discussion of the possible reason for differences in the results produced by the two analytical techniques.

*Application of Propensity Score Analysis*

Propensity score analysis has been conducted in various ways. The method applied in this paper is based on a procedure described by Yanovitzky et al. (2005). Yanovitzky et al. presented the following six steps researchers could use 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 subgroups of participants. In this illustration, three covariates were identified and labeled cov_1, cov_2, and cov_3.

*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 or a chi-square test of differences in proportions can be employed. 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 Score Group Formation. Before propensity group formation, statistically significant difference existed between the cov_1 means of the control and experimental groups. The probability values of the t tests used to statistically test the differences between means of the control and experimental groups with respect to cov_2 (p = .24) and cov_3 (p = .13) exceeded the .05 alpha level.

**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).

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 stepwise fashion. The stepwise 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
stepwise method used to enter the two-way interaction terms) is not the only method that can be used.

Once the stepwise procedure was completed the final procedure used to construct the logistic regression 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, cov_3, cov_1-by-cov_3, and cov_2-by-cov_3, are listed in Table 3.

____________________________________

Insert Table 3 about here

____________________________________

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. The final 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 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--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.

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 Score Group Formation. Recall that the column in Table 2 entitled Pre-Propensity Score 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 is judged to have reduced the initial bias on the covariates.

**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:

\[
\hat{\delta} = \frac{1}{N} \sum_{k=1}^{4} \frac{n_k}{N} (\bar{Y}_{ek} - \bar{Y}_{ck})
\]  

(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:

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

(Equation 2)

Where:

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

The execution of Step 6 of the propensity score analysis of the criterion scores produced the results listed in Table 4. Note that the tests were conducted as one-tailed tests because the mean of the experimental group was hypothesized to exceed the mean of the control group.

None of the independent $t$ tests of the differences between the criterion means of the control and experimental groups in Propensity Score Groups 1 through 3 produced $p$ values less than the one-tailed alpha level of .05. However, the $t$ test of the difference between the criterion means of the experimental (71.04) and control (66.22) for Propensity Score Group 4 ($t = 2.36, p = .01$) was statistically significant.

---

Insert Table 4 about here

---

Since the difference between the criterion means of the control and experimental groups was statistically significant for only one of the propensity score groups, it would not be appropriate to test the overall criterion mean difference between the control and experimental groups. The results suggest that the difference between the criterion means of the control and experimental groups exists for certain types of students but not other types of students. Thus, a *global interaction effect*, that is, an interaction effect in which the covariates are considered simultaneously, exits. When considering this global interaction effect, the researchers would conclude that the criterion scores will be higher for the students in the experimental group than the control group only for those students ...
whose covariate scores would place them in Propensity Score Group 4. For a discussion of how researchers could identify which future students would be placed in Propensity Score Group 4 see Fraas et al. (2006).

Application of ANCOVA

It would be typical for researchers to assess whether interaction effects between the treatments and the three covariates were related to the criterion scores by conducting the analysis as follows:

1. An initial multiple regression model is constructed that consists of the following seven predictor 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.

2. If a coefficient for any of the interaction variables was not statistically significant, it would be deleted from the model, that is, the amount of variation accounted for by the variable would be pooled in the error term.

3. Any statistically significant interaction term would be plotted and interpreted. This type of application of ANCOVA, however, would not produce an assessment of the global interaction effect. Thus, a comparison of the results produced by this approach to ANCOVA and the results produced by the propensity score analysis would not be legitimate. Fraas et al. (2006) presented an initial discussion of how researchers could use multiple linear regression models to estimate a global interaction effect.

To produce ANCOVA results that would allow a legitimate comparison, that is, one that considers the global interaction effect, a researcher would complete eight steps.
**Step 1.** The first multiple linear regression model is constructed as follows:

\[
\text{criterion} = a_0 + b_1(\text{treatment}) + b_2(\text{cov}_1) + b_3(\text{cov}_2) + b_4(\text{cov}_3) + \epsilon_1 \quad \text{(MLR Model 1)}
\]

Where:

1. The symbols \(a_0\) and \(\epsilon_1\) 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.

The results of the analysis of this model are listed in Table 5.

---

Insert Table 5 about here

---

**Step 2.** Once the partial regression weights are obtained, a variable, which is labeled \(\text{com\_cov}\), is computed by summing the products of each covariate and its corresponding partial regression weight. That is, \(\text{com\_cov}\) is calculated as follows:

\[
\text{com\_cov} = b_2(\text{cov}_1) + b_3(\text{cov}_2) + b_4(\text{cov}_3) \quad \text{(Equation 3)}
\]

For this illustration, the values for \(b_2 (0.48480)\), \(b_3 (0.10272)\), and \(b_4 (0.01706)\) would be inserted into Equation 3.

**Step 3.** Once this new variable (\(\text{com\_cov}\)) is created, a second new variable, which is labeled \(\text{com\_cov-by-treatment}\), is created by multiplying the \(\text{com\_cov}\) variable by the treatment variable.

**Step 4.** After this interaction variable (\(\text{com\_cov-by-treatment}\)) is generated, a second multiple linear regression model is constructed and analyzed. In this model the
criterion scores are regressed onto the treatment, com_cov, and com_cov-by-treatment variables. Thus, this second model is as follows:

\[
\text{criterion} = a_0 + b_1(\text{treatment}) + b_2(\text{com_cov}) + b_3 (\text{com_cov -by-treatment}) + \varepsilon_2 \quad \text{(MLR Model 2)}
\]

The results of the analysis of MLR Model 2 are contained in Table 6.

-----------------------------

Insert Table 6 about here

-----------------------------

Step 5. A t test of the \( b_3 \) coefficient in MLR Model 2 is conducted to indicate whether the global interaction effect is statistically significant (see Table 6). The \( b_3 \) (0.265) is statistically significant \((t = 2.86, p < .01)\).

Step 6. Since this global interaction effect is statistically significant, the two regression lines produced by MLR Model 2 (one regression line for the control group and one regression line for the interaction effect) are calculated and plotted on a graph in which the Y and X axes are labeled Criterion Scores and Com_Cov Scores, respectively. To obtain the regression line for the control group, the value for the treatment variable is set equal to zero. Two values are selected for the com_cov variable, with one near the bottom and one near the top of the range of values. The values selected were 20 and 60. With the value for com_cov set equal to 20, the value for the criterion variable is calculated as follows for the control group using the parameter estimates for MLR Model 2:

\[
\text{criterion} = 27.641 - 8.219(\text{treatment}) + 0.890(\text{com_cov}) + 0.265(\text{com_cov-by-treatment})
\]
criterion = 27.641 - 8.219(0) + 0.890(20) + 0.265(20)(0)

criterion = 45.4

With com_cov set equal to 60, the value for the criterion variable is calculated as follows:

criterion = 27.641 - 8.219(0) + 0.890(60) + 0.265(60)(0)

criterion = 81.00

The two values needed to generate the line for the experimental group, which were obtained in the same manner as the values for the control group except the value for the treatment variable was set equal to 1, are 42.5 and 88.7. These four points are used to plot the interaction effect (see Figure 1).

**Step 7.** The interaction effect depicted in Figure 1 reveals the interaction effect is disordinal. Since the interaction effect is disordinal, the point of intersection of the two regression lines (one regression line for the control group and one regression line for the interaction effect) is calculated. This point of intersection is calculated by first, setting the value for the treatment variable equal to 0 in MLR Model 2. Second, the value for the treatment variable is set equal to 1. Third, these two terms are set equal to each other as follows:

\[
27.641 - 8.219(0) + 0.890(\text{com}\_\text{cov}) + 0.265(\text{com}\_\text{cov})(0)
\]

\[
= 27.641 - 8.219(1) + 0.890(\text{com}\_\text{cov}) + 0.265(\text{com}\_\text{cov})(1) \quad (\text{Equation 4})
\]

Fourth, this equation is solved for the value of com_cov, which is 31.0. Thus, the two regression lines intersect at the com_cov value of 31.0. An examination of the disordinal interaction indicates that students with com_cov scores below 31.0 have higher estimated criterion scores in the control group than in the experimental group, while students with
com_cov scores above 31.0 have higher estimated criterion scores in the experimental group than in the control group.

Step 8. The Johnson-Neyman confidence limits are calculated for the interaction effect (Johnson & Neyman, 1936). See Pedhazur (1973) for a discussion of this technique and Fraas and Newman (1997) for an SPSS® syntax file that can be used to calculate the limits. In the program provided by Fraas and Newman, the following 12 values are required to calculate the limits:

1. The sum of the com_cov squared values for the control group (5185.44).
2. The sum of the com_cov squared values for the experimental group (3716.85).
3. The number of students in the control group (123).
4. The number of students in the experimental group (129).
5. The residual sum of squares value for MLR Model 2 (4613.59).
6. The mean com_cov score for the control group (36.90).
7. The mean com_cov score for the experimental group (40.01).
8. The slope of the regression line for the control group (.890).
9. The slope of the regression line for the experimental group (1.155).
10. The y-intercept point of the regression line for the control group (27.641).
11. The y-intercept point of the regression line for the experimental group (19.422).
12. The critical $F$ value with $df_e = 4$ and $df_n = 248$ is 3.03. Twice this value (6.06) is used in the program to obtain simultaneous limits.

Substituting these values into the SPSS® syntax file produced limits of -18.6 and 36.6. Since the lower limit is below any meaningful value of com-cov, only the upper limit is plotted on the graph contained in Figure 1. With a lower limit of -18.6, none of the
students would be identified as “benefiting” from the control method. Only the students who have com_cov scores above 36.6 would be identified as “benefiting” from the experimental method.

Comparison of the Results of the Two Analytical Techniques

It is important for researchers to compare the results produced by the propensity score analysis and the ANCOVA. This assessment should compare various results produced by the two analytical procedures, including the following:

1. Global interaction effect—Both the propensity score analysis and the ANCOVA identified a statistically significant global interaction effect in which students with certain covariate values were estimated to have higher criterion scores when exposed to the experimental group. The propensity score analysis revealed that the mean criterion score for the experimental group was significantly higher than the mean criterion score for the control group for only one of the four propensity score groups (Propensity Score Group 4), while the Johnson-Neyman confidence limits for the disordinal interaction indicated that criterion scores for the experimental group were statistically significantly higher for students with com_cov scores above 36.6.

2. Crosstabulation of student classification—To further examine the degree of similarity of results produced by the two analytical techniques, a crosstabulation of the numbers of students who were and who were not identified as “benefiting” from exposure to the experimental method by each analytic technique should be conducted. It should be noted that to assess the degree of agreement in the crosstabulation, one must remember that the difference between the mean scores
of the control and experimental groups in Propensity Score Group 4 was

*statistically significant*. Thus, to generate a legitimate comparison of consistent
classification, the researchers need to construct Johnson-Neyman confidence
limits (Johnson & Neyman, 1936) for the statistical interaction effect revealed by
the ANCOVA. These limits will allowed us to identify the com_cov scores above
and below which the criterion scores were statistically significantly different for
subjects in the control and experimental groups.

A review of crosstabulation, which is listed in Table 7, reveals that 63
students were identified by the propensity score analysis as “benefiting” from
exposure to the experimental method (i.e., the students in Propensity Score Group
4), while the ANCOVA identified 171 such students (i.e., the students with
com_cov scores above 36.6). If the 63 students who were identified by the
propensity score analysis as “benefiting” from the experimental method, 60 were
also identified by ANCOVA as “benefiting” from the experimental method.
However, the ANCOVA identified an additional 111 students who would
“benefit” from exposure to the experimental group. In addition, only 138 (54.8%)
of the students were consistently classified by both analytic methods. The results
forced us to address the questions: *What is the reason for the lack of agreement
between the results produced by the two analytic techniques employed?*

Insert Table 7 about here
Possible Reason for the Difference in Results

One of the easiest ways to understand the conceptual differences between the results produced by propensity score analysis and ANCOVA is though the use of the Ballantine Approach as discussed by Cohen and Cohen (1983) and the use of two multiple linear regression models, which reflect the visual representation of the Ballantine Approach. This approach was used by Cohen and Cohen to examine the differences between semipartial and partial correlations. In a similar manner we will use this approach to clarify the nature of the criterion scores being analyzed by propensity score analysis and ANCOVA.

When ANCOVA is conducted, the variance in the criterion variable that is being accounted for by predictor variable is the adjusted proportion of variance in the criterion variable. As indicated in Figure 2, in ANCOVA the proportion of variance in the criterion variable uniquely accounted for by the predictor variable is represented only by area $A$. Thus, the proportion of variance in the criterion variable accounted for by areas $B + C$ has been removed. By comparison, when propensity score analysis is used, the groups are equated before the analysis of the criterion variable is conducted. Thus, areas $B$ and $C$ are not partitioned out (i.e., removed) from the criterion variable. Thus, the area equal to $A + B + C$ is included in the criterion variable when it is regressed onto the predictor variable. Thus, none of the variation in the criterion variable is removed from the analysis.

---------------------------------------------------------------

Insert Figure 2 about here

---------------------------------------------------------------
This concept can also be illustrated by the use of the following regression models, which reflect part correlation procedures:

\[ y = a_0 + b_1(\text{covariate variable}) + \text{res}_1 \] (MLR Model 3)

\[ \text{res}_1 = a_1 + b_1(\text{predictor variable}) + \text{res}_2 \] (MLR Model 4)

If the assumptions regarding the distribution of the residuals in MLR Model 3 are met and the correlation coefficient value for the covariate values and the criterion values is greater than zero and less than one, the residual values will be zero correlated with the predicted criterion values. In MLR Model 3, each residual value is the difference between its corresponding \( y \) value and its predicted \( y \) value. Therefore, in MLR Model 3, the residuals can be thought of as adjusted \( y \) scores. MLR Model 4 reflects a question that is conceptually similar to ANCOVA, which is referred to as part correlation. A review of MLR Model 4 reveals that it analyzes how well the predictor variable can predict adjusted criterion values that have been adjusted for by the covariates. This is conceptually the intent of ANCOVA. Since the residuals are composed of the adjusted \( y \) scores, the adjusted \( y \) scores are different from the \( y \) scores (i.e., the unadjusted \( y \) scores). Thus, ANCOVA may be analyzing a different construct than the original construct represented by the criterion variable.

This concept can be demonstrated through the use of a Ballantine Approach diagram. The variation in the adjusted \( y \) scores (residual values) is represented by the total area contained in the Criterion Variable circle minus the two areas of B and C, while the variation for the unadjusted \( y \) scores is represented by the total area contained in the Criterion Variable circle. Thus, the unadjusted scores and the adjusted \( y \) scores are not the same.
The question for researchers to consider is: Since these two sets of values, that is, the unadjusted scores and the adjusted scores represent different constructs, what is a good approach to analyze a difference in the criterion variable between groups when selection bias exists? Tracz et al. (2005) recommend that if researchers use ANCOVA to analyze the difference between adjusted means, they should establish the reliability and validity of the residualized or adjusted scores. Fraas et al. (2006) noted that since this would be no small task, researchers may find it more practical to utilize propensity analysis. In propensity score analysis, the issue of selection bias is addressed by analyzing unadjusted means in propensity score groups rather than the adjusted means, as is the case in ANCOVA.

Summary

Propensity score analysis and ANCOVA are two analytic techniques that can be used to analyze data obtained from non-randomized designs. These two techniques, however, in general, do not produce similar results. In the results presented in this paper, both techniques identified a disordinal global interaction effect between the covariates and the treatment variable. However, the numbers of students identified as “benefiting” from the experimental method by the two analytic methods varied considerably.

One possible explanation for this difference is the nature of the criterion variable analyzed by each analytic technique. The propensity score analysis attempts to equate the treatment groups by identifying propensity score groups who have similar covariate values for the control and experimental groups. Once these propensity score groups are formed, the differences between the unadjusted criterion means of the control and experimental groups are analyzed for each propensity group. In ANCOVA, the variation
in the criterion scores of the control and experimental groups are analyzed once they have been *adjusted for the covariate values*. Thus, the ANCOVA analyzes *adjusted* criterion mean scores.

As can be seen through the use of the Ballantine Approach and the multiple linear regression models, the propensity score analysis and the ANCOVA analyze different criterion variables and, hence, different constructs. If these two analytic approaches do indeed analyze different constructs, they will not be addressing the same research question. Thus, it is incumbent on the researchers to ensure that the appropriate analytic technique is used to address the specific research question of interest.
Reference


Table 1

*Descriptive Statistics for the Criterion Variable and the Covariates*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Control</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Criterion</td>
<td>60.47</td>
<td>7.03</td>
</tr>
<tr>
<td>Cov_1</td>
<td>23.76</td>
<td>8.29</td>
</tr>
<tr>
<td>Cov_2</td>
<td>227.82</td>
<td>26.49</td>
</tr>
<tr>
<td>Cov_3</td>
<td>116.02</td>
<td>13.89</td>
</tr>
</tbody>
</table>

*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

Pre- and Post-Propensity Score Group Formation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pre-Propensity score group formation</th>
<th>Post-Propensity score group formation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p$</td>
<td>$p$</td>
</tr>
<tr>
<td>Cov_1</td>
<td>&lt;.01</td>
<td>.41</td>
</tr>
<tr>
<td>Cov_2</td>
<td>.24</td>
<td>.66</td>
</tr>
<tr>
<td>Cov_3</td>
<td>.13</td>
<td>.90</td>
</tr>
</tbody>
</table>
Table 3

*Results for the Final Logistic Regression Model used to Generate the Propensity Scores*\(^a\)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Wald test value</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cov_1</td>
<td>-0.417</td>
<td>2.21</td>
<td>.14</td>
</tr>
<tr>
<td>Cov_2</td>
<td>0.255</td>
<td>7.72</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Cov_3</td>
<td>0.306</td>
<td>5.81</td>
<td>.02</td>
</tr>
<tr>
<td>Cov_1-by- Cov_3</td>
<td>0.005</td>
<td>4.14</td>
<td>.04</td>
</tr>
<tr>
<td>Cov_2-by- Cov_3</td>
<td>-0.002</td>
<td>8.21</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Constant</td>
<td>-38.375</td>
<td>6.77</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

\(^a\)\(\Delta(-2\text{Log likelihood value}) = 68.995, \chi^2 = 98.96, df = 5, p < .01,\) Cox and Snell \(R^2 = .24,\)

Nagelkerke \(R^2 = .32.\)
### Table 4

*Estimated Treatment Effects on Criterion Scores Using Propensity Score Groups*

<table>
<thead>
<tr>
<th>Propensity Score groups</th>
<th>Treatment</th>
<th>Group size</th>
<th>M (SD)</th>
<th>t</th>
<th>p&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>Control</td>
<td>51</td>
<td>57.86 (8.72)</td>
<td>0.40</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>12</td>
<td>58.92 (5.43)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>36</td>
<td>60.72 (4.91)</td>
<td>0.04</td>
<td>.48</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>27</td>
<td>60.78 (7.34)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 2</td>
<td>Control</td>
<td>27</td>
<td>63.15 (3.96)</td>
<td>0.25</td>
<td>.40</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>36</td>
<td>63.47 (5.80)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 3</td>
<td>Control</td>
<td>9</td>
<td>66.22 (4.18)</td>
<td>2.36</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>54</td>
<td>71.04 (5.87)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 4</td>
<td>Control</td>
<td>123</td>
<td>61.99&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>129</td>
<td>63.55&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> One-tailed p values

<sup>b</sup> 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 5

Results of MLR Model Constructed to Generate the Regression Coefficients used to

Compute the Com_Cov Variable (MLR Model 1)\(^a\)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>(t)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>2.04796</td>
<td>3.29</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Cov_1</td>
<td>0.48480</td>
<td>9.99</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Cov_2</td>
<td>0.10272</td>
<td>6.21</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Cov_3</td>
<td>0.01706</td>
<td>0.58</td>
<td>.56</td>
</tr>
<tr>
<td>Constant</td>
<td>23.57321</td>
<td>7.45</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

\(^a\)\(R^2 = .690, df_n = 4, df_d = 247, F = 137.456, p < .001, R^2 = .685\)
Table 6

Results of MLR Model used to Estimate and Statistically Test the Global Interaction

Effect (MLR Model 2) \(^a\)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>(t)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment (^b)</td>
<td>-8.219</td>
<td>-2.26</td>
<td>.02</td>
</tr>
<tr>
<td>Com_Cov</td>
<td>0.890</td>
<td>14.86</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Com_Cov-by-Treatment</td>
<td>0.265</td>
<td>2.86</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Constant</td>
<td>27.641</td>
<td>12.32</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

\(^a\) \(R^2 = .700, df_n = 3, df_d = 248, F = 192.82, p < .001, \overline{R^2} = .696\)
Table 7

*Crosstabulation of Students Identified as the Type Who Would Benefit from Exposure to the Experimental Method.*

<table>
<thead>
<tr>
<th>Classification of students with ANCOVA</th>
<th>Would not benefit</th>
<th>Would benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification of students with propensity score analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Would not benefit</td>
<td>78&lt;sup&gt;a&lt;/sup&gt;</td>
<td>111</td>
</tr>
<tr>
<td>Would benefit</td>
<td>3</td>
<td>60&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup>A total of 138 students (78 + 60) received the same classification by both methods (54.8%).
Figure 1. Global interaction effect
Figure 2. Ballantine approach to variation accounted for in the criterion variable