2016 ■ Vol. 12 ■ no. 3

Violation of the homogeneity of regression slopes assumption in ANCOVA for two-group pre-post designs: Tutorial on a modified Johnson-Neyman procedure

Teresa R. Johnson[a] ✉

[a]Johns Hopkins University School of Medicine, Baltimore

Abstract ■ Aptitude-treatment interaction (ATI) effects are present when individuals demonstrate differential outcomes across treatments based upon aptitude—that is, any measurable individual characteristic, attribute, or ability (e.g., anxiety, learning style, motivation, prior knowledge). ATI effects may exist in data from one design commonly used in psychological and educational research—the two-group pre-post design—in which pre-intervention scores may be considered to reflect individual aptitude. Researchers may mistakenly overlook these effects, however, due to inappropriate analytical approaches. When applying analysis of covariance (ANCOVA), it is important to check for ANCOVA assumptions, including an assumption known as homogeneity of regression slopes. When heterogeneity of regression slopes is found, ATI effects are revealed. Consequently, alternative approaches to ANCOVA must be sought. Using formulae based on the Johnson-Neyman procedure to define simultaneous regions of significance is one straightforward alternative. This tutorial outlines the process for analyzing data resulting from two-group pre-post studies when data violate the ANCOVA assumption of homogeneity of regression slopes. What was initially viewed as an obstacle may result in the discovery of an ATI effect, which may be described statistically through simple mathematical calculations.

Keywords ■ Aptitude-treatment interaction effects; two-group pre-post designs; ANCOVA; Johnson-Neyman procedure. **Tools** ■ SPSS.

✉ Teresa.Johnson@jhmi.edu

TRJ: 0000-0001-7468-6947

10.20982/tqmp.12.3.p253

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer.

Introduction

Pre-post designs are used extensively in psychological and educational research to assess changes in outcomes between two time points and/or to compare outcomes of independent groups. Of particular interest here, are two-group pre-post designs. While these research designs may seem straightforward, analysis of data resulting from such studies often presents great challenges (e.g., Lord, 1967).

Common analytical approaches for examining data from two-group pre-post designs include (1) independent samples t-test on gain or difference scores, (2) mixed model analysis of variance (ANOVA), often called repeated-measures analysis of variance (RM-ANOVA) with a between-subjects factor, and (3) analysis of covariance (ANCOVA). When choosing an approach, it is important to understand the strengths, limitations, and requirements of each method; generally speaking, however, ANCOVA is usually the preferred approach (Bonate, 2000; Dimitrov & Rumrill, 2003; Dugard & Todman, 1995; Huck & McLean, 1975; Knapp & Schafer, 2009; Overall, 1993; Senn, 1994).

This tutorial provides guidance for students and researchers who originally planned to use ANCOVA for the analysis of data arising from two-group pre-post studies, who exercised due diligence in checking to see if the data met assumptions required for ANCOVA, and who now find themselves at an impasse because the data violated an assumption of ANCOVA known as homogeneity of regression

slopes. For those with little to moderate training in statistics and/or who are consulting textbooks for guidance, the coursework and text likely concluded with the directive that ANCOVA should not be employed under these circumstances. What are proposed solutions, then? Fortunately, what was initially viewed as an obstacle may result in the discovery of what is known as an aptitude-treatment interaction (ATI) effect. Moreover, ATI effects may be described statistically through a series of simple mathematical calculations. Example analyses and results are presented using SPSS 23.0 (IBM; Armonk, NY); figures depicting scatterplots were created using Excel 2013. This tutorial assumes that the reader has a basic understanding of the use of SPSS.

Understand ATI effects and the importance of their identification

ATI effects are present when individuals demonstrate differential performance across treatments based upon aptitude. "Performance" refers to a measurable outcome of interest, "treatment" represents a manipulated exposure variable (e.g., psychological or educational intervention), and "aptitude" denotes a measurable learner characteristic, attribute, or ability (e.g., anxiety, learning style, motivation, prior knowledge).

For example, consider a hypothetical investigation of learners' scores on an achievement test (performance) following participation in either a high- or low-structure educational intervention (treatment) based upon learners' scores on a measure of trait anxiety (aptitude). An interaction would be observed if low-anxiety learners performed better on the test following participation in low-structure education as compared to high-structure education, whereas high-anxiety learners performed better on the test following participation in high-structure education as compared to low-structure education. In this scenario, the optimal level of structure to define in the educational environment for achieving the outcome goals depends upon (or interacts with) a learner's level of trait anxiety. The concept of ATI effects aligns well with other areas of focus in psychological and educational research and practice, such as theories of multiple intelligences, individual differences, motivation, differentiated instruction, and learning style.

ATI effects were recognized as points of inquiry in the field of psychology several decades ago (Cronbach, 1957), although some methods for identifying and handling these types of interactions were proposed as early as the 1930s (e.g., P. O. Johnson & Neyman, 1936; also see Cronbach & Snow, 1969). Resources for expanding one's understanding of ATI effects, including their history and significance, and for conducting ATI research are widely accessible (e.g., Bracht, 1970; Cronbach & Snow, 1969; Driscoll, 1987; Snow, 1989, 1991; Snow, Federico, & Montague, 1980), and are recommended as a first step toward identifying and handling ATI effects in one's own research. The importance to researchers and educators is apparent because, "In general, unless one treatment is clearly best for everyone, treatments should be differentiated in such a way as to maximize their interactions with aptitude variables" (Cronbach, 1957, p. 681). ATI research helps us to discover how this differentiation should be approached.

Although the steps in this tutorial are intentionally situated within the context of the two-group pre-post design, note that research studies may be designed *a priori* specifically to detect ATI effects, and may examine numerous treatments, outcomes, and aptitudes (e.g., Abelson, 1953; P. O. Johnson & Hoyt, 1947; Karpman, 1983; Potthoff, 1964).

Check for missing data, outliers, and basic assumptions

As with all ANOVA models, data must first be screened to assess the degree to which there are any missing data and/or outliers, and to confirm that data meet basic assumptions of ANOVA, including normality and homogeneity of variance. Statistics textbooks that address ANOVA models are great resources for furthering one's understanding of these initial exploratory steps and for assisting in making decisions related to handling missing data, outliers, non-normality, and heterogeneity of variance (e.g., Roberts & Russo, 1999; Tabachnick & Fidell, 2012; Turner & Thayer, 2001). Additionally, numerous resources are available to students and researchers that provide step-by-step instructions for creating boxplots, frequency distributions, and other graphical displays of data, conducting statistical tests of assumptions (e.g., Shapiro-Wilk's test of normality; Levene's test of homogeneity of variance), and performing other related tasks in a statistical analysis program (e.g., Cunningham & Aldrich, 2011; Field, 2013, for SPSS 18.0 and SPSS 20.0-21.0 respectively). Let's assume here that missing data and outliers are either non-existent or are handled appropriately, and that assumptions of normality and homogeneity of variance are either met or the researcher considers ANCOVA to be robust to slight violations here. ANCOVA models, in particular, have two additional assumptions that must be met—linearity and homogeneity of regression slopes. Due to the distinct roles they play in ANCOVA models, these assumptions are addressed individually in the next two sections.

View scatterplots and check for linearity

Following ANCOVA models and their application to two-group pre-post designs, treat post-test score as the dependent variable (y), treatment group as the independent variable (or between-subjects factor; using "Treatment 1" and "Treatment 2" to denote groups), and pre-test score as the

Table 1 ■ Hypothetical Data Set for Two-Group Pre-Post Design, Including Squared Pre-Test Scores and Summary Statistics ($N = 40$)

Pre-Test (x)		Pre-Test Squared (x^2)		Post-Test (y)	
Treatment 1 ($n_1 = 20$)	Treatment 2 ($n_2 = 20$)	Treatment 1 ($n_1 = 20$)	Treatment 2 ($n_2 = 20$)	Treatment 1 ($n_1 = 20$)	Treatment 2 ($n_2 = 20$)
20	25	400	625	49	78
26	26	676	676	54	84
32	32	1024	1024	51	77
37	36	1369	1296	63	81
41	45	1681	2025	61	82
42	46	1764	2116	74	87
44	46	1936	2116	67	84
47	49	2209	2401	70	89
47	51	2209	2601	72	89
52	55	2704	3025	74	86
54	55	2916	3025	85	84
54	59	2916	3481	81	86
57	62	3249	3844	73	82
61	63	3721	3969	90	86
65	67	4225	4489	97	85
68	68	4624	4624	95	83
68	71	4624	5041	97	85
70	72	4900	5184	98	87
75	76	5625	5776	100	83
79	82	6241	6724	99	86
$\bar{x}_1 = 51.95$ $SD_{x_1} = 16.28$	$\bar{x}_2 = 54.30$ $SD_{x_2} = 16.37$	$\sum x_1^2$ = 59013	$\sum x_2^2$ = 64062	$\bar{y}_1 = 77.50$ $SD_{y_1} = 16.97$	$\bar{y}_2 = 84.20$ $SD_{y_2} = 3.16$

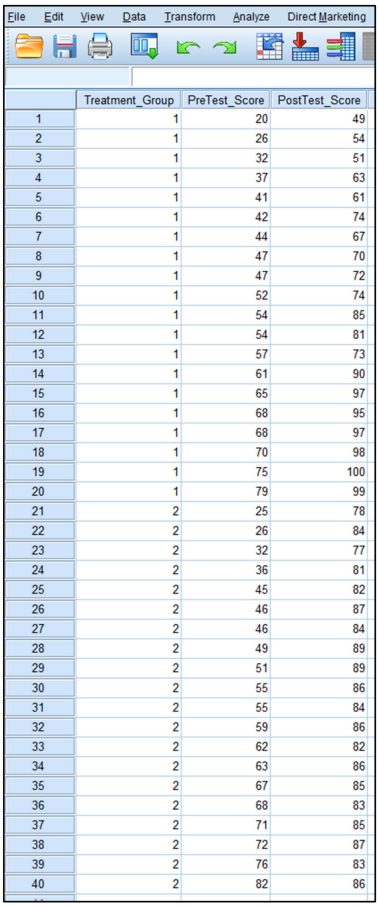
covariate (x). A hypothetical data set of 40 scores is provided in Table 1 to facilitate a fully-worked concrete example across the remaining steps. A screenshot of the data arranged in SPSS appears in Figure 1. Scores were defined deliberately to emphasize patterns and outcomes, and are ordered in the table from low to high based on pre-test score for ease of review.

ANCOVA assumes linearity—that is, that there is a straight-line relationship between the covariate and dependent variable. Inspection of bivariate scatterplots is appropriate for fundamental assessment of linearity (residuals plots may be used for additional diagnoses). Pre- and post-test scores in the hypothetical data set, indeed, demonstrate a linear relationship, as shown in Figure 2, Panel (A); as such, Pearson's r may be used to describe the magnitude of this relationship. The correlation here is 0.72, which is significantly different than 0 ($p < 0.001$). However, a scatterplot depicting a single regression line for the entire sample may mask additional characteristics of the data.

To explore further and to prepare for testing the homogeneity of regression slopes assumption, construct the scatterplot to depict regression lines for each treatment, as shown in Figure 2, Panel (B). Linear relationships are observed here for both treatments ($r = 0.96, p < 0.001$ for Treatment 1; $r = 0.46, p = 0.04$ for Treatment 2); however, the slopes (angles) of the lines appear very different. The fact that the lines are not parallel indicates an interaction; moreover, the fact that the lines cross signifies what is known as a disordinal interaction, wherein the superiority of one treatment over the other is not constant across the full range of pre-test scores (Aiken & West, 1991; Cronbach & Snow, 1981; Lubin, 1961).

Although a researcher would construct the two-group scatterplot for the purposes of testing homogeneity of regression slopes anyway, note that careful examination of these hypothetical data at earlier stages should have already caught the researcher's attention. Standard deviation (SD) for pre-test scores between treatments is nearly identical (16.3 vs. 16.4), but the discrepancy in SD for post-test scores between treatments is marked (16.9 vs. 3.2; which, incidentally, resulted in a violation of homogeneity of variance; Levene's test, $p < 0.001$). The significantly smaller variation in post-test scores for Treatment 2 implies that the regression line for that group would be comparatively flatter (i.e., relative to the x-axis) than the re-

Figure 1 ■ Screenshot of the hypothetical data set for a two-group pre-post design in SPSS 23.0 ($N = 40$).



	Treatment_Group	PreTest_Score	PostTest_Score
1	1	20	49
2	1	26	54
3	1	32	51
4	1	37	63
5	1	41	61
6	1	42	74
7	1	44	67
8	1	47	70
9	1	47	72
10	1	52	74
11	1	54	85
12	1	54	81
13	1	57	73
14	1	61	90
15	1	65	97
16	1	68	95
17	1	68	97
18	1	70	98
19	1	75	100
20	1	79	99
21	2	25	78
22	2	26	84
23	2	32	77
24	2	36	81
25	2	45	82
26	2	46	87
27	2	46	84
28	2	49	89
29	2	51	89
30	2	55	86
31	2	55	84
32	2	59	86
33	2	62	82
34	2	63	86
35	2	67	85
36	2	68	83
37	2	71	85
38	2	72	87
39	2	76	83
40	2	82	86

gression line for Treatment 1 (note that if all students in Treatment 2 achieved the same post-test score, where SD = 0, all points would lie perfectly on a straight flat line parallel to the x-axis).

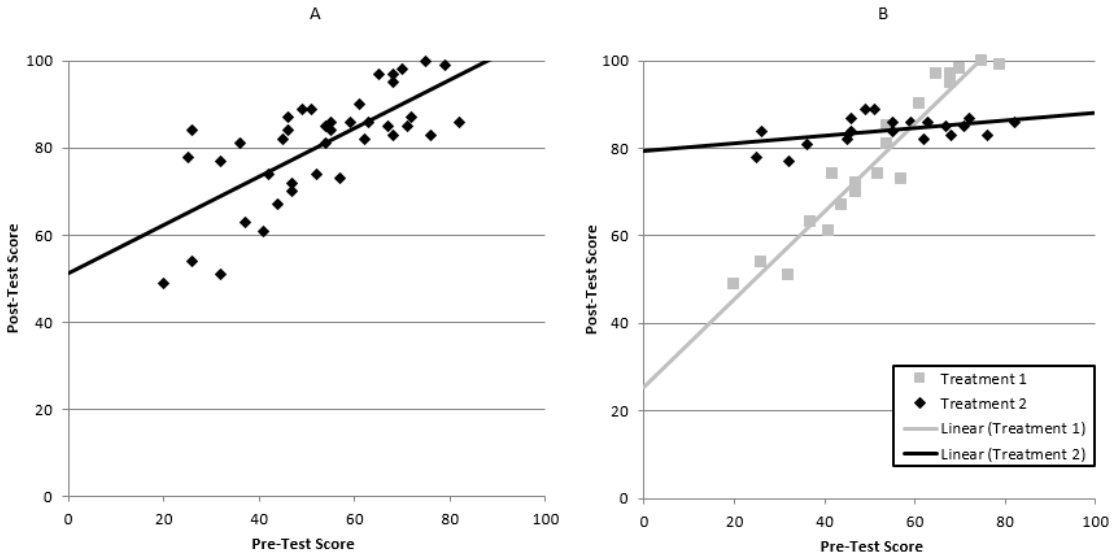
Check for homogeneity of regression slopes

At this point, an examination of the raw data set and related descriptive statistics, and observation of the non-parallel, crossing lines in the two-group scatterplot suggest the presence of an ATI effect and possibly render ANCOVA an inappropriate analytical approach. However, the extent to which the lines are non-parallel needs to be formally tested as a check for homogeneity of regression slopes.

To test whether or not the slopes of the regression lines for the two groups are equal, enter all terms into a univariate general linear model—that is, enter the independent variable (treatment), the covariate (pre-test score), and an independent variable-by-covariate (treatment-by-pre-test score) interaction term, using post-test score as the dependent variable. There are two ways to accomplish this from an analysis standpoint. SPSS, for example, performs this test under Analyze >General Linear Model>Univariate, provided the researcher specifies a custom model to force the inclusion of all three terms, as follows:

1. Move `PostTest_Score` to the "Dependent Variable" box, move `Treatment_Group` to the "Fixed Factor(s)" box, and move `PreTest_Score` to the "Covariate(s)" box;
2. Click the "Model" button;
3. Select the "Custom" radio button;
4. Ensure that the "Build Term(s)" type displays "Interaction" (as this is the default when accessing this dialog box for the first time), highlight both `Treatment_Group` and `PreTest_Score`, and move the two variables over to the "Model" window;
5. Change the "Build Term(s)" type to "Main effects," highlight both `Treatment_Group` and `PreTest_Score`, and move the two variables over

Figure 2 ■ Two pre-/post-test score scatterplots. Panel (A) shows the pre-/post-test score scatterplot for all participants ($N = 40$), and Panel (B) shows the pre-/post-test score scatterplot for participants by treatment ($n_1 = 20$; $n_2 = 20$).



to the "Model" window;

6. Click "Continue," and then click "OK."

Figure 3 shows the two relevant dialog boxes in SPSS properly filled.

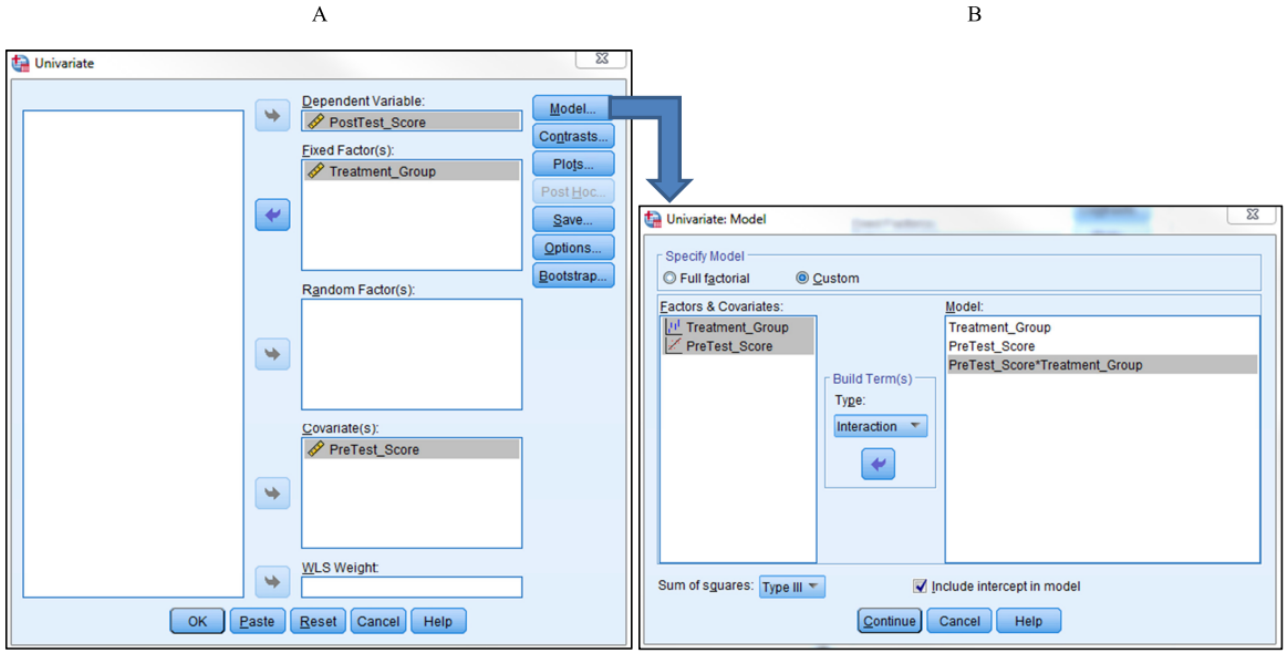
The test may also be conducted through linear regression analysis in SPSS (i.e., Analyze>Regression>Linear), provided the researcher first creates a new variable—the treatment-by-pre-test score interaction term—for inclusion in the model (e.g., Transform>Compute Variable; name the new "Target Variable," and define `Treatment_Group*PreTest_Score` as the "Numeric Expression").

If the p-value corresponding to the interaction term is less than the pre-specified Type I error rate (e.g., $\alpha = 0.05$), then the slopes are considered non-equivalent and the homogeneity of regression slopes assumption of ANCOVA is violated. Tables 2 and 3 display results derived from SPSS using the hypothetical data set, and testing redundantly for homogeneity of regression slopes through the SPSS general linear model and regression approaches, respectively. The two approaches are formally identical—$t^2 = 11.95^2 = 142.80$ is the same as $F = 142.64$, rounding errors notwithstanding. Either analysis formally confirms the presence of a significant ATI effect for this hypothetical data set, and necessitates an alternative analytical approach.

Identify an alternative analytical approach

The presence of an interaction effect suggests that treatment effects are not the same across various levels of the co-variate or aptitude. Proceeding with ANCOVA, despite heterogeneity of regression slopes, may lead researchers to erroneously conclude that performance did not differ significantly by treatment. Consequently, alternative analytical approaches must be considered, which may include the Johnson-Neyman procedure (Fraas & Newman, 1997; Karpman, 1983; Kowalski, Schneiderman, & Willis, 1994; Rogosa, 1981) and extensions thereof, such as that proposed by Potthoff (1964). For the sake of simplicity, let's consider formulae constructed by Potthoff (1964) as a modification to the Johnson-Neyman procedure (Aiken & West, 1991; D'Alonzo, 2004; Pedhazur & Schmelkin, 1991; Rogosa, 1981). These formulae allow for calculations of the point of intersection (crossover point) of regression lines, and what are known as simultaneous regions of significance (SROS). Potthoff (1964) explains that a "simultaneous" region of significance is a region that "with confidence ≥ 95 percent (for $\alpha = .05$), we can state that the two groups [...]are different simultaneously for all points contained in it" (p. 244). Regarding the hypothetical data set, SROS may be calculated to identify the pre-test score ranges for which treatments differ significantly on the post-test. The purpose of this alternative approach is to obtain values associated with two SROS (R')—one region wherein post-test scores for Treatment 1 are significantly higher than post-

Figure 3 ■ Screenshots of dialog boxes for testing the homogeneity of regression slopes assumption in SPSS 23.0. Panel (A) shows the dialog box used for defining variables, and Panel (B) shows the dialog box used for specifying the model, which is accessed by clicking the "Model" button in the previous dialog box, as identified by the start of the blue arrow in Panel (A).



test scores for Treatment 2, and one region wherein post-test scores for Treatment 2 are significantly higher than post-test scores for Treatment 1, as illustrated by equation (1). Derivations of each of the terms in the formulae below are described in subsequent sections.

Simultaneous Regions of Significance (R')

$$R' = \frac{-B \pm \sqrt{B^2 - AC}}{A} \tag{1}$$

where A, B, and C are defined as follows:

$$A = \frac{-2F_\alpha}{N-4}(ss_{res})\left[\frac{1}{\sum x_1^2} + \frac{1}{\sum x_2^2}\right] + (b_1 - b_2)^2 \tag{2a}$$

$$B = \frac{2F_\alpha}{N-4}(ss_{res})\left[\frac{\bar{x}_1}{\sum x_1^2} + \frac{\bar{x}_2}{\sum x_2^2}\right] + (a_1 - a_2)(b_1 - b_2) \tag{2b}$$

$$C = \frac{-2F_\alpha}{N-4}(ss_{res})\left[\frac{N}{n_1 n_2} + \frac{\bar{x}_1^2}{\sum x_1^2} + \frac{\bar{x}_2^2}{\sum x_2^2}\right] + (a_1 - a_2)^2 \tag{2c}$$

Crossover Point (P)

$$P = \frac{a_1 - a_2}{b_2 - b_1} \tag{3}$$

in which $\bar{x}_1$ and $\bar{x}_2$ are the aptitude means for treatments 1 and 2, respectively; $\sum x_1^2$ and $\sum x_2^2$ are the sum of squares of the aptitude for treatments 1 and 2, respectively; ss_{res} is the residual sum of squares from the overall regression analysis when all terms of the design are included; a_1 and a_2 are the intercepts for treatments 1 and 2, respectively; b_1 and b_2 are the regression coefficients for treatments 1 and 2, respectively; N is the total number of participants; n_1 and n_2 are the number of participants in treatments 1 and 2, respectively; and F_α is the tabled value of F with 2 and $N-4$ degrees of freedom at a pre-determined α level.

Methods for obtaining these values are as follows:

$\bar{x}_1$, $\bar{x}_2$, $\sum x_1^2$, and $\sum x_2^2$. The notation $\bar{x}_1$ and $\bar{x}_2$ indicates the mean of x for treatments 1 and 2, respectively. The notation $\sum x_1^2$ and $\sum x_2^2$ symbolizes the sum of all squared x values for treatments 1 and 2, respectively. In the case of the hypothetical data set, x is the aptitude or pre-test score, and the two groups are Treatment 1 and Treatment 2. These values are presented in Table 1.

ss_{res}. The notation ss_{res} represents the residual (error) sum of squares from the overall regression analysis when all terms of the design are included. Fortunately, an analysis appropriate to obtain this value was already completed

Table 2 ■ Test for Homogeneity of Regression Slopes Using a General Linear Model in SPSS 23.0

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*-value
Corrected Model	5573.32	3	1857.77	124.83	< 0.001
Intercept	9019.96	1	9019.96	606.07	< 0.001
Treatment	2406.09	1	2406.09	161.67	< 0.001
Pre-Test Score	3028.40	1	3028.40	203.49	< 0.001
Treatment-by-Pre-Test Score	2123.53	1	2123.53	142.69	< 0.001[a]
Error	535.78[b]	36	14.88		
Total	267578.00	40			

Note. [a]: p-value corresponding to interaction term; if $p < \alpha$, homogeneity of regression slopes assumption is violated; [b]:ss_{res}.

Table 3 ■ Test for Homogeneity of Regression Slopes Using Regression Analysis in SPSS 23.0

Model	Sum of Squares	*df*	Mean Square	*F*	*p*-value
Regression	5573.32	3	1857.77	124.83	< 0.001
Residual	535.78[b]	36	14.88		
Total	6109.10	39			

Model	Coefficient	Standard Error	*t*	*p*-value
Constant	-28.76	6.65	-4.32	< 0.001
Treatment	54.07	4.25	12.72	< 0.001
Pre-Test Score	1.92	0.12	15.82	< 0.001
Treatment-by-Pre-Test Score	-0.92	0.08	-11.95	< 0.001[a]

Note. [a]: p-value corresponding to interaction term; if $p < \alpha$, homogeneity of regression slopes assumption is violated; [b]: ss_{res}

earlier—that is, when testing for homogeneity of regression slopes through univariate general linear model analysis. Recall that "all terms" of the design included the independent variable (treatment), covariate (pre-test score), and independent variable-by-covariate (treatment-by-pre-test score) interaction term, using post-test score as the dependent variable. SPSS, for example, displays this value as Error Type III Sum of Squares following the general linear model approach, as shown in Table 2, and as Residual Sum of Squares following the regression approach, as shown in Table 3.

a_1, a_2, b_1, and b_2. The notation a_1 and a_2 correspond to the intercepts from linear regression analyses for treatments 1 and 2, respectively; b_1 and b_2 represent the regression coefficients for treatments 1 and 2, respectively. These values, as shown in Table 4, may be obtained by conducting separate linear regression analyses by treatment, using pre-test score as the independent variable and post-test score as the dependent variable. In SPSS, for example, filter for Treatment 1 cases (i.e., Data>Select Cases; write "If condition is satisfied" statement, "Treatment = 1") and then access the regression analysis through Analyze>Regression>Linear, as shown in Figure 4; repeat the process filtering for Treatment 2 cases.

N, n_1, n_2, and F_α Document N (total number of participants), and n_1 and n_2 (number of participants in treatments 1 and 2, respectively). Next, recall that F_α is the tabled value of F with 2 and $N - 4$ degrees of freedom (df) at a pre-determined α level. These values may be looked up in F tables typically appearing in the appendices of statistics textbooks or online at various locations (e.g., search in a web browser for "table of critical F values"). Ensure that the table being viewed corresponds to the α level specified by the researcher (e.g., $\alpha = 0.05$). To find the appropriate F value in the table, note that the value of 2 represents the numerator df, whereas the value of $N - 4$ represents the denominator df.

Finally, complete all calculations using the values obtained in the preceding steps. Values derived from the hypothetical data set are as follows:

Figure 4 ■ Screenshot of the dialog box for performing linear regression analysis in SPSS 23.0.

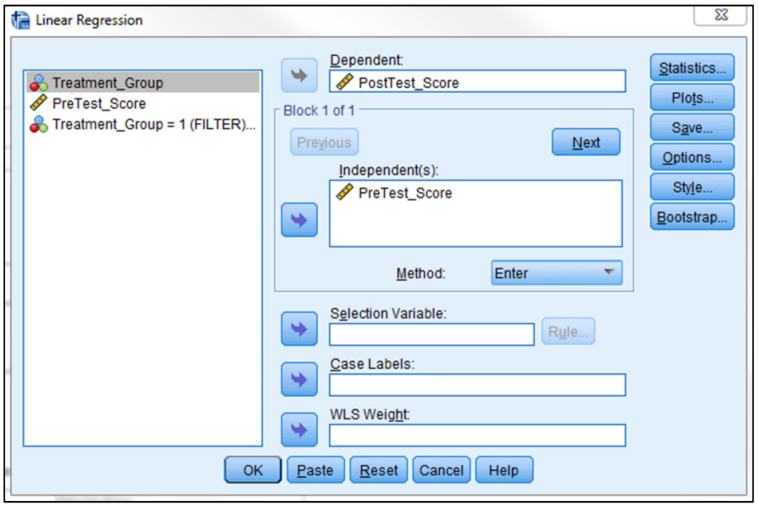


Table 4 ■ Derivation of a_1, a_2, b_1, and b_2 using separate regression analyses by treatment group in SPSS 23.0

Model	Coefficient	Standard Error	t	p-value
Treatment 1				
Constant	25.31^a	3.55	7.13	< 0.001
Pre-Test Score	1.01^b	0.07	15.38	< 0.001
Treatment 2				
Constant	79.37^a	2.28	34.79	< 0.001
Pre-Test Score	0.09^b	0.04	2.21	0.04

Note. ^{a}: Value represents intercept (a); ^{b}: Value represents regression coefficient (b)

$$\bar{x}_1 = 51.95 \text{ and } \bar{x}_2 = 54.30;$$
$$\sum x_1^2 = 59013 \text{ and } \sum x_2^2 = 64062;$$
$$ss_{res} = 535.78;$$
$$a_1 = 25.31 \text{ and } a_2 = 79.37;$$
$$b_1 = 1.01 \text{ and } b_2 = 0.09;$$
$$N = 40;$$
$$n_1 = 20 \text{ and } n_2 = 20;$$
$$F_{0.05}(2,36) = 3.26.$$

Substituting these values in the Potthoff (1964) formulae presented earlier, it may be found that $P = 58.8$ and $R' = 55.4$ and 62.2, the temporary variables A, B, and C having values of 0.8432, -49.5675, and 2903.876, respectively. Therefore, the crossover point of the two regression lines is 58.8, the value associated with the lower SROS is 55.4, and the value associated with the upper SROS is 62.2.

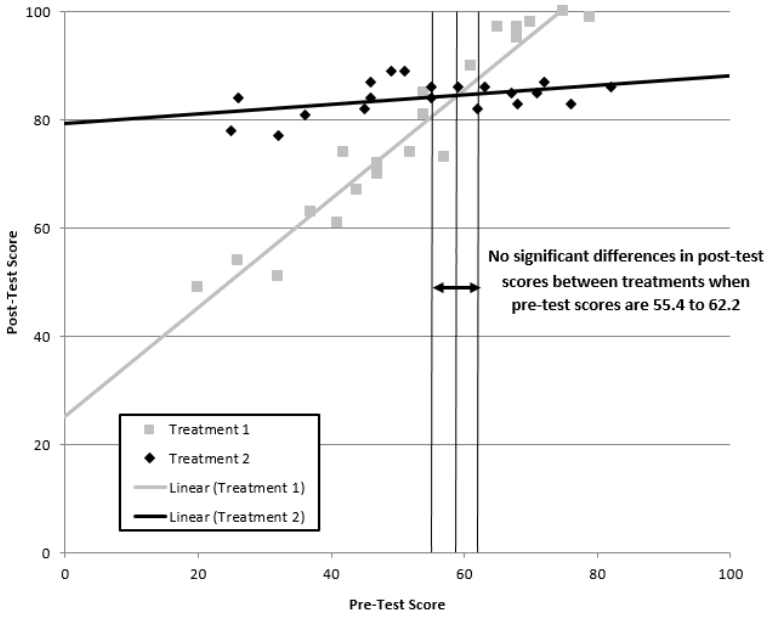
Re-construct two-group scatterplot to depict and interpret SROS and crossover point

It is helpful to view the SROS and crossover point within the context of the full data set. The two-group scatterplot originally constructed to informally assess homogeneity of regression slopes is well-suited for this purpose. The insertion of shapes and text boxes facilitates the depiction of the calculations rather easily, as shown in Figure 5.

Interpretations may be best explained by leveraging the hypothetical data set and related results. In general, the presence of an ATI effect indicates that treatments had a differential effect on students' post-test performance depending upon student aptitude, or pre-test score (Cronbach & Snow, 1981; Pedhazur & Schmelkin, 1991).

Specifically, calculations completed at this point may be interpreted as follows, as displayed in Figure 5: (1) at pre-test scores below 58.8 (crossover point), students in Treatment 2 tended to score higher on the post-test than did students in Treatment 1; (2) at pre-test scores above

Figure 5 ■ Pre-/post-test score scatterplot by treatment with SROS and crossover point noted.



58.8, students in Treatment 1 tended to score higher on the post-test than did students in Treatment 2; (3) at pre-test scores below 55.4 (value of lower SROS), students in Treatment 2 scored significantly higher on the post-test than did students in Treatment 1; (4) at pre-test scores above 62.2 (value of upper SROS), students in Treatment 1 scored significantly higher on the post-test than did students in Treatment 2; and (5) at pre-test scores occurring in the range of 55.4-62.2 (also called the "region of insignificance," D'Alonzo, 2004, p. 808), post-test scores did not differ significantly between treatments.

Explore outcomes between treatments in each region of significance

At this stage, researchers and educators may be additionally interested in the magnitude of differences between treatments occurring in each region of significance, as "statistically significant" does not always equate to practically important or educationally meaningful. Depending upon sample size within each region, a review of descriptive statistics, mean comparisons, and confidence intervals (CI) of mean differences may be helpful.

For example, regarding the hypothetical data set, 23 students scored below 55.4 on the pre-test—12 students in Treatment 1 (post-test mean score = 66.8; SD = 11.5), and 11 students in Treatment 2 (post-test mean score = 83.7; SD = 4.0). The mean difference was 16.9 (95%CI = [9.4, 24.6], equal variances not assumed, df = 13.84). Thirteen students scored above 62.2 on the pre-test—6 students in Treatment 1 (post-test mean score = 97.7; SD = 1.8), and 7 students in Treatment 2 (post-test mean score = 85.0; SD = 1.5). The mean difference was 12.7 (95%CI = [10.7, 14.7], equal variances assumed, df = 11). Confidence intervals for these hypothetical data were obtained through independent samples t-tests in SPSS.

Discoveries of ATI effects necessitate consideration of generalizability and, in the context of two-group pre-post designs, future use of the pre-test (e.g., to identify optimal treatment assignment). The nature of ATI effects—indicating that individuals perform differently under different conditions depending upon aptitude—highlights the need to determine resource availability (e.g., professionals, space, equipment, and assessment for the delivery and evaluation of tailored treatments or interventions). Differentiation of treatments, interventions, instruction, and other services may place greater demands upon the provider, but the benefits to recipients may make tailoring the experiences worthwhile.

Conclusion

Two-group pre-post designs may seem simplistic in their structure; however, the analytical approach to handling data arising from these designs is critical. Approaches that are commonly applied to two-group pre-post designs may

yield markedly different results and interpretations depending upon the nature of the data. For example, returning to the hypothetical data set, (1) if the researcher opted not to control for pre-test scores (since means were not significantly different between the two groups, $p = 0.65$), an independent samples t-test on post-test scores indicated no significant between-group differences ($p = 0.10$, equal variances not assumed); (2) an independent samples t-test on the gain or difference scores indicated no significant between-group differences (p = 0.23, equal variances not assumed); (3) mixed RM-ANOVA indicated a significant main effect of time ($p < 0.001$), but no significant main effect of treatment ($p = 0.29$), or time-by-treatment interaction effect ($p = 0.23$); and (4) ANCOVA indicated no significant differences between treatments after controlling for pre-test scores ($p = 0.05$). Results from each of these analyses would leave the researcher concluding that performance does not differ significantly between treatments; moreover, perhaps observation of the low power associated with ANCOVA (0.50) would have led the researcher to instead increase the sample size for a future study trial.

It is important to check for ANCOVA assumptions, and when heterogeneity of regression slopes occurs, an ATI effect has been discovered. Accordingly, an alternative approach to ANCOVA must be sought. Formulae provided by Potthoff (1964) as a modification to the Johnson-Neyman procedure (Aiken & West, 1991; D'Alonzo, 2004; Pedhazur & Schmelkin, 1991; Rogosa, 1981) is one straightforward alternative.

While these steps were expressly formulated to apply only to two-group pre-post designs, note that methods have been suggested for handling data from designs that incorporate more than two groups (e.g., Potthoff, 1964), more than one covariate (e.g., Karpman, 1983), and more than one outcome (e.g., Potthoff, 1964).

References

Abelson, R. P. (1953). A note on the Neyman-Johnson technique. *Psychometrika*, *18*(3), 213–218. doi:10 . 1007 / bf02289058

Aiken, L. S. & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage Publications.

Bonate, P. L. (2000). *Analysis of pretest-posttest designs*. Boca Raton, FL: Chapman & Hall/CRC.

Bracht, G. H. (1970). Experimental factors related to aptitude-treatment interactions. *Review of Educational Research*, *40*(5), 627–645. doi:http://dx.doi.org/10.3102/00346543040005627

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671–684. doi:10.1037/h0043943

Cronbach, L. J. & Snow, R. E. (1969). *Final report: Individual differences in learning ability as a function of instructional variables*. Stanford, CA: Stanford University Press.

Cronbach, L. J. & Snow, R. E. (1981). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington Publishers.

Cunningham, J. B. & Aldrich, J. O. (2011). *Using SPSS: An interactive hands-on approach*. Thousand Oaks, CA: Sage Publications.

D'Alonzo, K. T. (2004). The Johnson-Neyman procedure as an alternative to ANCOVA. *Western Journal of Nursing Research*, *26*(7), 804–812. doi:10 . 1177 / 0193945904266733

Dimitrov, D. M. & Rumrill, P. D. (2003). Pretest-posttest designs and measurement of change. *Work*, *20*(2), 159–165.

Driscoll, M. P. (1987). Aptitude-treatment interaction research revisited. In *Paper presented at the annual convention of the association for educational communications and technology, february*. Atlanta, GA, USA.

Dugard, P. & Todman, J. (1995). Analysis of pre-test post-test control group designs in educational research. *Educational Psychology*, *15*(2), 181–198. doi:10 . 1080 / 0144341950150207

Field, A. (2013). *Discovering statistics using IBM SPSS statistics (4th ed.)* Thousand Oaks, CA: Sage Publications.

Fraas, J. W. & Newman, I. (1997). The use of the Johnson-Neyman confidence bands and multiple regression models to investigate interaction effects: Important tools for educational researchers and program evaluators. *Multiple Linear Regression Viewpoints*, *24*, 14–24.

Huck, S. W. & McLean, R. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: a potentially confusing task. *Psychological Bulletin*, *82*(4), 511–518. doi:10.1037/h0076767

Johnson, P. O. & Hoyt, C. (1947). On determining three dimensional regions of significance. *Journal of Experimental Education*, *15*(3), 203–212. doi:10 . 1080 / 00220973.1947.11010357

Johnson, P. O. & Neyman, J. (1936). Tests of certain linear hypotheses and their applications to some educational problems. *Statistical Research Memoirs*, *1*, 57–63.

Karpman, M. B. (1983). The Johnson-Neyman technique using SPSS or BMDP. *Educational and Psychological Measurement*, *43*(1), 137–147. doi:10 . 1177 / 001316448304300117

Knapp, T. R. & Schafer, W. D. (2009). From gain score t to ANCOVA F (and vice versa). *Practical Assessment, Research & Evaluation*, *14*(6), 1–7.

Kowalski, C. J., Schneiderman, E. D., & Willis, S. M. (1994). ANCOVA for nonparallel slopes: the Johnson-Neyman technique. *International Journal of Bio-Medical Computing*, *37*(3), 273–286. doi:10.1016/0020-7101(94)90125-2

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*(5), 304–305. doi:10.1037/h0025105

Lubin, A. (1961). The interpretation of significant interaction. *Educational and Psychological Measurement*, *21*(4), 807–817. doi:10.1177/001316446102100406

Overall, J. E. (1993). Letter to the editor: The use of inadequate corrections for baseline imbalance remains a serious problem. *Journal of Biopharmaceutical Statistics*, *3*(2), 271–276. doi:10.1080/10543409308835066

Pedhazur, E. J. & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Potthoff, R. F. (1964). On the Johnson-Neyman technique and some extensions thereof. *Psychometrika*, *29*(3), 241–256. doi:10.1007/BF02289721

Roberts, M. J. & Russo, R. (1999). *A student's guide to analysis of variance*. New York: Routledge.

Rogosa, D. (1981). On the relationship between the Johnson-Neyman region of significance and statistical tests of parallel within-group regressions. *Educational and Psychological Measurement*, *41*(1), 73–84. doi:10.1177/001316448104100108

Senn, S. (1994). Testing for baseline balance in clinical trials. *Statistics in Medicine*, *13*(17), 1715–1726. doi:10.1002/sim.4780131703

Snow, R. E. (1989). Aptitude-treatment interaction as a framework for research on individual differences in learning. In L. Ackerman, R. J. Sternberg, & R. R. Glaser (Eds.), *P* (pp. 13–59). New York: Learning, individual differences: Advances in theory, and research . : W.H. Freeman.

Snow, R. E. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of Consulting and Clinical Psychology*, *59*(2), 205–216. doi:10.1037/0022-006X.59.2.205

Snow, R. E., Federico, P., & Montague, W. (1980). *Aptitude, learning, and instruction (vols. 1 & 2)*. Hillsdale, NJ: Erlbaum.

Tabachnick, B. G. & Fidell, L. S. (2012). *Using multivariate statistics (6th ed.)* Boston, MA: Pearson/Allyn & Bacon.

Turner, J. R. & Thayer, J. F. (2001). *Introduction to analysis of variance: design, analysis, & interpretation*. Thousand Oaks, CA: Sage Publications.

Citation

Johnson, T. R. (2016). Violation of the homogeneity of regression slopes assumption in ANCOVA for two-group pre-post designs: Tutorial on a modified Johnson-Neyman procedure. *The Quantitative Methods for Psychology*, *12*(3), 253–263. doi:10.20982/tqmp.12.3.p253

Received: 13/04/2016 ~ Accepted: 10/06/2016