European Journal of Training and Development

Analyzing data from a pretest-posttest control group design: The importance of statistical assumptions

Linda Zientek Kim Nimon Bryn Hammack-Brown

Received 6 August 2015
Revised 4 February 2016
Accepted 16 March 2016

Analyzing data from a pretest-posttest control group design

The importance of statistical assumptions

Linda Zientek
Sam Houston State University, Huntsville, Texas, USA, and
Kim Nimon and Bryn Hammack-Brown
The University of Texas at Tyler, Tyler, Texas, USA

Abstract

Purpose – Among the gold standards in human resource development (HRD) research are studies that test theoretically developed hypotheses and use experimental designs. A somewhat typical experimental design would involve collecting pretest and posttest data on individuals assigned to a control or experimental group. Data from such a design that considered if training made a difference in knowledge, skills or attitudes, for example, could help advance practice. Using simulated datasets, situated in the example of a scenario-planning intervention, this paper aims to show that choosing a data analysis path that does not consider the associated assumptions can misrepresent findings and resulting conclusions. A review of HRD articles in a select set of journals indicated that some researchers reporting on pretest-posttest designs with two groups were not reporting associated statistical assumptions and reported results from repeated-measures analysis of variance that are considered of minimal utility.

Design/methodology/approach – Using heuristic datasets, situated in the example of a scenario-planning intervention, this paper will show that choosing a data analysis path that does not consider the associated assumptions can misrepresent findings and resulting conclusions. Journals in the HRD field that conducted pretest-posttest control group designs were coded.

Findings – The authors' illustrations provide evidence for the importance of testing assumptions and the need for researchers to consider alternate analyses when assumptions fail, particularly the homogeneity of regression slopes assumption.

Originality/value – This paper provides guidance to researchers faced with analyzing data from a pretest-posttest control group experimental design, so that they may select the most parsimonious solution that honors the ecological validity of the data.

Keywords Experiment, Regression, Assumptions, Homogeneity of regression slopes, Quantitative methods, Statistical tests

Paper type Conceptual paper

For several decades, researchers in the human resource development (HRD) field have been interested in determining if an intervention has made a change in knowledge, skills and/or attitudes. As noted by Russ-Eft and Hoover (2005, p. 94), "experimental and quasi-experimental designs can help advance HRD by aiding researchers and practitioners to determine cause-and-effect relationships". While there are several design options, we focus on the pretest-posttest control group design. Analyzing data

European Journal of Training and Development
Vol. 40 No. 8/9, 2016
pp. 638-659

2046-9012
DOI 10.1108/EJTD-08-2015-0066

from a pretest-posttest control group design can seem daunting and confusing, as researches have multiple data analysis strategies from which to choose (Huck and McLean, 1975). When researchers choose an inappropriate test or do not consider associated assumptions, the results may misrepresent data. Relevant to the pretest-posttest control group design, our goals are to review and illustrate the importance of considering statistical assumptions, and benchmark how HRD researchers are reporting related analyses.

Review of analytic choices

When analyzing data from a pretest-posttest control group design, researchers make several analytic decisions that depend on the research question and how data meet associated statistical assumptions (Tables I and II). Presuming interval data, choices for dependent variables include posttest, gain (i.e. simple difference) or residualized scores. We focus on posttest and gain scores. Readers considering residualized scores should consult relevant literature (Forbes and Carlin, 2005; Kisbu-Sakarya *et al.*, 2013; Nimon and Henson, 2015; Rogosa *et al.*, 1982; Zumbo, 1999).

In regards to analysis, the *t*-test is the simplest analytic choice, as it compares differences in either posttest scores or gain scores by group (i.e. experimental, control). Analysis of covariance (ANCOVA) allows researchers to determine the effect of the intervention on posttest or gain scores that is not predictable from the pretest. Multiple linear regression (MLR), which subsumes the roles of *t*-test and ANCOVA (Zientek and Thompson, 2009), also allows researchers to isolate the effect of the intervention on posttest or gain scores and can be used under conditions when data do not meet the unique statistical assumption associated with ANCOVA (i.e. homogeneity of regression slopes; Gliner *et al.*, 2003; Linn and Slinde, 1977).

	Dependent variable	
Analysis	Post	Gain
t-test	Which group differed more on the posttest scores?	"What is the effect of the treatment on the change from pretest to posttest?" (Knapp and Schafer, 2009, p. 2)
ANCOVA	"What is the effect of the treatment on the posttest that is not predictable from the pretest (i.e., conditional on the pretest)?" (Knapp and Schafer, 2009, p. 2) "Would the groups have been different on the postmeasure if they had been equivalent on the covariate?" (Maxwell and DeLaney, 2004, p. 401)	What is the effect of the treatment on simple differences "that is not predictable from the pretest (i.e., conditional on the pretest)?" (cf. Knapp and Schafer, 2009, p. 2)
MLR[a]	How much of the variance in posttest scores was uniquely accounted for by group?	How much of the variance in simple difference scores was uniquely accounted for by group?

Notes: ANCOVA = analysis of covariance; MLR = multiple linear regression; [a]multiple linear regression can also be used to answer research questions associated with ANCOVA

Table I.
Research questions by analysis and dependent variable

Table II.
Statistical assumptions by analysis and dependent variable

Analysis	Dependent variable: Posttest	Dependent variable: Gain
t-test	Homogeneity of variance	Homogeneity of variance
	Pretest equivalence	Homogeneity of regression slopes
ANCOVA[a]	Covariate and independent variable uncorrelated	
	Covariate and dependent variable highly correlated and linearly related	
	Homogeneity of regression slopes	
	Homoscedasticity of residuals	
MLR[a]	Homoscedasticity of residuals	

Notes: ANCOVA = analysis of covariance, MLR = multiple linear regression; normality, independence of observations and samples and reliability are assumptions for *t*-test, ANCOVA and MLR; linearity is required for ANCOVA and MLR; [a]statistical assumptions for posttest and gain scores are the same

T-*test*

When using the *t*-test to analyze group differences in posttest or gains scores, homogeneity of variance is assumed as well as normality, independence of observations and reliability. For posttest scores, pretest equivalence (no group differences on pretest scores) is an additional assumption. Although pretest equivalence may be assumed in the case of a randomized design, pretest differences between groups should be checked before conducting a *t*-test on posttest scores (Huck and McLean, 1975; Humphreys, 1976). For gain scores, a *t*-test may be considered as an analytic strategy as long as the data meet the necessary assumptions, including homogeneity of regression slopes, which is an assumption often not considered when analyzing gain scores (Edwards, 1960).

Difference scores as the dependent variable have been criticized because of the conclusion that the difference scores will be unreliable (Cronbach and Furby, 1970; Linn and Slinde, 1977; Lord, 1963). Overall and Woodward (1975, p. 86) stated that while unreliability stemming from summation of measurement errors for difference scores might "be a problem for certain types of correlational studies, it is not a cause for concern in the use of simple difference scores to measure treatment-induced change in experimental research". Furthermore, others have demonstrated that it might not always be the case that difference scores are unreliable (Rogosa *et al.*, 1982; Williams and Zimmerman, 1996; Zimmerman and Williams, 1982). This discussion has spanned decades, which illustrates the complexity associated with difference scores. As noted by Overall and Woodward (1975), reliability of difference scores decreases as pretest-posttest correlation increases, and other things constant, a maximized test statistic value will be obtained when difference scores reliability equals zero. Thus, Overall and Woodward (1975, p. 86) concluded:

> […] the reliability of the original prescores and postscores is a valid concern, but this is not true of the *decrease* in reliability resulting from combining of measurement errors in the testing of group difference scores.

For large sample sizes, the Central Limit Theorem assures us that the sampling distribution of the sample means is approximately normal (Thompson, 2006a; Williams *et al.*, 2013). For small sample sizes, the assumption for a *t*-test is that the

dependent variable is drawn from a normal distribution. However, the *t*-test is somewhat robust to violations of normality and homogeneity of variance, especially when sample sizes are equal or near equal (Boneau, 1960). So much so that Boneau (1960, p. 63) noted that "the *t*-test is seen to be functionally nonparametric or distribution-free". As with all analyses in the general linear model, *t*-tests might produce invalid results if observations are not independent or data are unreliable. In the case of the former, researchers risk misestimating the effects of predictors on an outcome and reporting finding that are opposite to what might be reported when data are analyzed with an analytical method that honors the clustered nature of the data (Lane *et al.*, 2012; Osborne, 2000). In the case of the latter, effects are likely to be attenuated in the case of unreliable data, but may be inflated in the case of correlated error (Nimon *et al.*, 2012).

Of course, no analysis is complete without considering practical significance (Kirk, 1996). For analyses based on *t*-tests, standardized mean differences are typical effect sizes reported where the difference in the dependent variable is the numerator and the denominator is either the *standard deviation (SD)* for the control group (Glass's Delta; Kirk, 1996) or the pooled SD for the two groups (Cohen's *d*; Cohen, 1988). Effect size benchmarks of small (0.2), medium (0.5) and large (0.8) were established by Cohen (1988, p. 532) "with much diffidence, qualifications, and invitations not to employ them if possible".

Analysis of covariance

In the presence of group differences on pretests, ANCOVAs can be used to analyze group differences in posttest or gains scores. Note that ANCOVA with posttest scores produces identical results as ANCOVA with gain scores; thus, there is no advantage to conducting ANCOVAs with gain scores (Jamieson, 2004). The use of ANCOVAs is ill-advised for intact groups (e.g. gender groups) or groups that have not been randomly assigned, as it unlikely that the data will meet the associated statistical assumptions for analysis of variance (ANOVA) (Henson, 1998).

In addition to independence of observations and reliability, assumptions for ANCOVAs include homoscedasticity of residuals and homogeneity of regression slopes. Furthermore, the covariate and the independent variable should be uncorrelated, whereas the covariate and the dependent variable should be highly correlated and have a linear relationship (Miller and Chapman, 2001). As stated by Maxwell and Delany (2004), "in the population, the error scores ε_{ij}, must be independently and normally distributed" and "have an expected value of zero and a constant variance" (p. 421). When the homogeneity of regression slopes assumption is met, ANCOVA is robust to violations of normality (Levy, 1980). Failure to meet the homogeneity of regression slopes can yield misleading ANCOVA results (Henson, 1998; Huck and McLean, 1975). Even though in a balanced design with normal data, ANCOVA *F* appears to be robust against violations of the assumption of homogeneity of regression slopes, and caution should be given to the degree of robustness based on violations under various conditions (Hamilton, 1977; Harwell, 2003; Levy, 1980; Wu, 1984). In fact, Hamilton (1977, p. 712) went so far as to caution readers from generalizing beyond the situations investigated in his study and stated that:

> […] whether or not the results observed in this [his] study will hold for other slope combinations, other group sizes, more than two groups, etc. will have to await further research.

Research on violation of the assumption of homogeneity of regression slopes has focused on F statistics and not on effect sizes, the latter of which is a measure of practical significance. For ANCOVA, practical significance typically is measured with partial eta-squared, η_p^2 (Henson, 1998; Pedhazur, 1997). In ANCOVA, η_p^2 indicates how much group membership accounts for variance in the dependent variable after eliminating variance associated with pretest differences from both the grouping and dependent variables (Maxwell *et al.*, 1985; Miller and Chapman, 2001). Note, however that some researchers caution against the use of ANCOVA because "statistical corrections remove parts of the dependent variable" and the "covariance corrections may result in the analysis of a dependent variable that no longer makes any sense" (Thompson, 2006a, p. 356).

Multiple linear regression

Researchers can use MLR to analyze group differences in posttest or gain scores. An advantage of using MLR over ANCOVA is that group differences in the regression slopes between posttest and pretests scores can be specifically modeled by including an interaction term between the pretest and grouping variables. The independent variables will be the following: group, pretest scores and the interaction between the pretest and group (Gliner *et al.*, 2003; Linn and Slinde, 1977). As with ANCOVA, MLR produces identical similar results with gain and posttest scores. Analyzing either dependent variable yields the same test statistic and partial eta-squared (η_p^2) for the group effect. Although the squared semi-partial correlation coefficient (i.e. uniqueness coefficient) that indicates how much variance in the dependent variable is uniquely associated with group membership is different between posttest and gain scores when pretest differences exist, the incremental $\hat{y}$ values resulting from analyzing the posttest and gain scores are perfectly correlated.

Williams *et al.* (2013) reported that assumptions for MLR include reliability, linearity and errors that are homoscedastic, independent and normally distributed with a mean of zero. They also noted failure to meet homoscedasticity of residuals can result in untrusty results, and as sample sizes become larger, the normality assumption becomes less important. As previously noted, data should also have independence of observations to yield valid results (Osborne, 2000).

Conducting a commonality analysis on the resulting regression effect allows the researcher to compute a η_p^2 as in ANCOVA. Alternatively, researchers may use the unique commonality coefficient associated with group to determine much variance in the dependent variable is uniquely associated with the intervention (Kraha *et al.*, 2012).

Purpose of the study

Determining if interventions make an improvement is important to improving practices in the field of HRD. When analyzing data from a pretest-posttest control group design, researchers must consider associated assumptions of statistical tests and choose the best test that fit the data. Otherwise, findings may misestimate the effect of an intervention. The purposes of the present study are to illustrate that analyses might produce different results depending on whether certain assumptions are met, and provide benchmarks of current HRD practices for analyzing data from pretest-posttest control group designs.

Illustrations

To illustrate how failure to address key statistical assumptions can make a difference in the interpretation of data from a pretest-posttest control group design, we present two examples based on simulated data. In the first example, we analyzed posttest scores as the dependent variable. In the second example, we analyzed gain scores as the dependent variable.

Posttest scores

Using R syntax provided in Appendix 1, we generated a simulated dataset using descriptive statistics reported in Chermack *et al.* (2015) to illustrate how the magnitude of an intervention depends on the analytical approach chosen and why it is important to match the analytic approach to the associated assumptions. In Chermack *et al.* (2015), participants in a scenario-planning workshop ($n = 48$) as well as a control group ($n = 42$) completed pre- and post-surveys containing items from the Situational Outlook Questionnaire (SOQ). As indicated by Isaksen and Akkermans (2011, p. 170), "the SOQ is an online questionnaire consisting of 53 closed-ended questions on a four-point Likert scale […] [analyzing] […] the creation of an organizational climate that supports innovation". In Chermack *et al.* (2015), data were collected on a five-point Likert scale, and composite scores were created for each scale. Thus, the data considered were interval in nature (Norman, 2010). The SOQ yields nine factors. Pertinent to our illustration is the pretest-posttest data on play/humor, as there were significant differences in pretest scores and the data did not meet the assumption of homogeneity of regression slopes.

We analyzed the playfulness/humor posttest scores using four different techniques. First, we used the posttest score as the dependent variable and group (i.e. 0 = control and 1 = intervention) as the independent variable with both a *t*-test and an ANOVA. As seen in Table III, both analyses resulted in similar results, as expected. The effect sizes indicate that the workshop had a positive effect on posttest scores, where posttest scores were 0.78 of a *SD* higher for the intervention group and that group explained 13.43 per cent of the variance in posttest scores[1]. Note, however, that these results would only be appropriate to report if the groups' pretest scores were equivalent, which they were not ($M_{CTL} = 2.55$, $M_{INT} = 3.02$; Chermack *et al.*, 2015).

Next we used ANCOVA as a means to control for pretest differences between groups. The results indicate that after controlling for pretest differences, the intervention accounted for 9.46 per cent of the residual variance. Also note that ANCOVA produced adjusted means that were slightly different than the observed pretest and posttest

Analysis	Control *M*	Intervention *M*	*p*-value for group	Effect size type	Effect size (%)
t-test	2.56	3.06	0.0003825	d	0.78
ANOVA	2.56	3.06	0.0003825	η^2	13.43
ANCOVA	2.59[a]	3.03[a]	0.0033760	η_p^2	9.46
Regression	2.46[a]	2.96[a]	0.0005902	η_p^2	12.87
				sr^2_{Group}	11.16

Notes: [a] Adjusted means; sr^2_{Group} = uniqueness coefficient

Table III. Comparison of simulated Dataset 1 results by analysis type of posttest scores

scores. However, these results would only be appropriate to report if the correlation between the pretest and posttest was the same between the two groups. As depicted in Figure 1, the assumption of homogeneity of regression slopes was clearly not met. While the correlation between pretest and posttest scores was positive for the intervention group ($r = 0.41$), it was negative for the control group ($r = -0.29$).

Finally, we modeled the data using regression, where we examined the group effect taking into account pretest differences that were allowed to vary by group. The homoscedasticity of residuals assumption was met. This analysis was followed up with a commonality analysis which indicated that group accounted for 12.87 per cent of the variance in the residualized dependent variable and 11.16 per cent of the variance in posttest scores. It can also be seen in Table III that the means from the regression equation are slightly different than the observed means from *t*-test or adjusted means from ANCOVA.

While only the regression results that modeled posttest scores using pretest scores, group membership and the interaction between pretest scores and group membership would be appropriate to report, readers can see how results might have been misinterpreted had a different analytic strategy been conducted that ignored the related statistical assumptions. Although group was a statistically significant factor in all models, reporting that the intervention accounted for 13.43 per cent of the variance in posttest scores would be misleading in the presence of pretest differences, as some of that explained variance was also attributable to pretest differences as well as the interaction between pretest differences and group membership. As well, the η_p^2 from

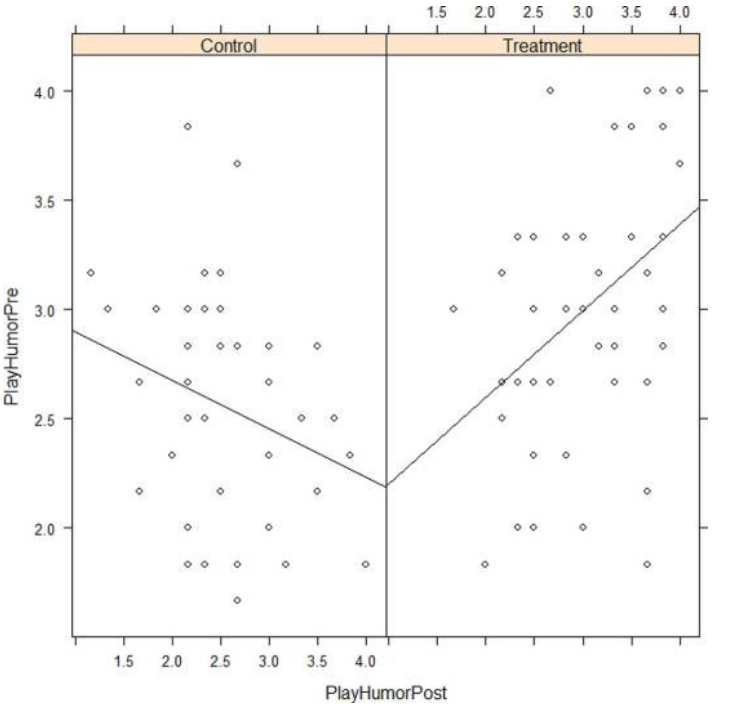

Figure 1.
Regression slope of pretest and posttests scores by group

ANCOVA also represented explained variance that was common with the interaction between pretest differences and group membership.

Simple difference versus posttest scores

Using R syntax provided in Appendix 2, we generated a simulated dataset ($n_1 = n_2 = 40$) to illustrate that analyses of gain scores will produce the same results as posttest scores analyses when the assumption that there are no pretest differences between groups is met. Pretest means for both groups were simulated to be 3.0. Posttest means for the control and experimental group were simulated to be 3.0 and 4.0, respectively. Standard deviations were set to 1 for all variables. As well, the data were simulated to meet the assumption of homogeneity of regression slopes, so as to validly create the simple difference scores. Correlations between pretest and posttest scores were simulated to be 0.5 in both the experimental and control groups (see Figure 2 for a scatterplot of data).

We ran the same set of analyses that we conducted in the first example. First, we ran *t*-test and ANOVA on the simple difference scores. Next, we conducted an ANCOVA on the difference scores using the pretest as a covariate. Then, we conducted an MLR of the difference scores using the pretest and the interaction between the pretest and the grouping variable as predictors. Finally, we replicated the aforementioned analyses using posttests scores as the dependent variable.

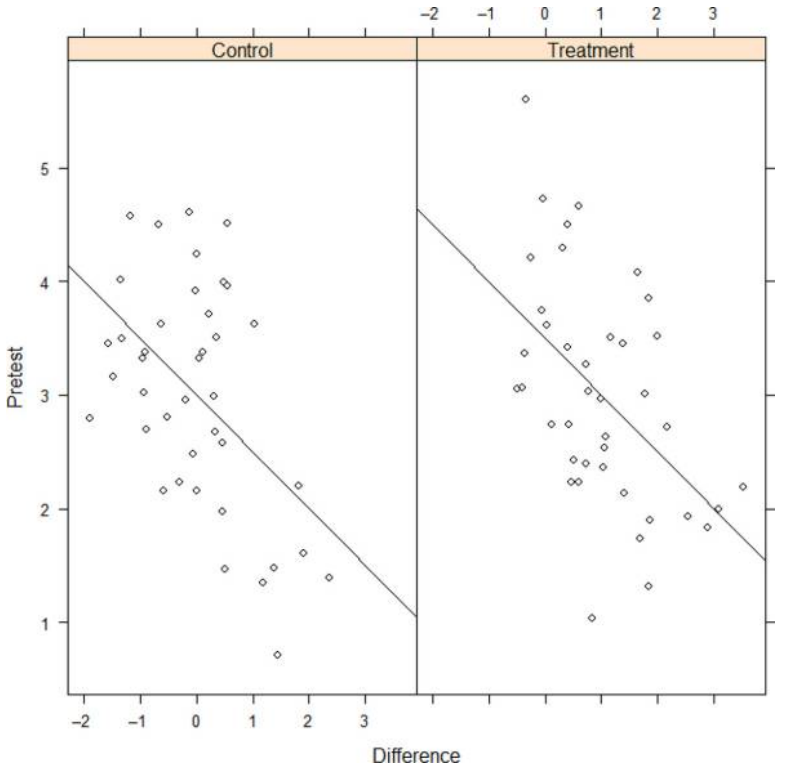

Figure 2.
Regression slope of pretest and posttest scores of Dataset 2 by group

As seen in Tables IV and V, analyzing difference scores and posttest scores produced identical results because mean pretest scores for the control and intervention groups were simulated to be identical. Also see that the *t*-test, ANOVA and regression results yielded the same measures of magnitude. Both the ANOVA and regression results indicated that group accounted for 20.41 per cent of the variability in difference (and posttests) scores. Note that the *d* of 1.00 is comparable to an η^2 of 20.41 per cent[1]. The only difference across the analyses is that the effect size for the ANCOVA is a partial η^2 and indicates that after eliminating variance associated with pretest differences in the grouping and dependent variables, 25.48 per cent of the remaining variance was shared.

Journal analyses

To establish a benchmark of relevant reporting practices, we conducted a five-year review of research in HRD-related journals from 2010 to 2014. The journals coded were *Advances in Developing Human Resources (ADHR)*, *European Journal of Training and Development*, *Human Resource Development International, Human Resource Development Quarterly* and *International Journal of Human Resources Development and Management*. We conducted an in-text search for "experiment" in all journals except *ADHR*, where text search was not available. Therefore, for *ADHR*, a search for "experiment" was conducted within every article that was not classified by the journal as a preface, future or next steps article, which resulted in a search of 142 *ADHR* articles. If author(s) identified the study as an experiment, we reported the authors' description of the study design and coded the group assignment as random or non-random. For studies with pretest-posttest design with two groups, the statistical test, descriptive statistics and associated assumptions reported were coded.

Of the 692 articles reviewed, 22 studies were identified by authors as an experiment and eight of those were specified as random assignment. More specifically, authors

Table IV.
Comparison of simulated Dataset 2 results by analysis type of simple difference scores

Analysis	Control *M*	Intervention *M*	*p*-value for group	Effect size type	Effect size (%)
t-test	0.00	1.00	2.596e-05	*d*	1.00
ANOVA	0.00	1.00	2.596e-05	η^2	20.41
ANCOVA	0.00[a]	1.00	2.596e-05	η_p^2	25.48
Regression	0.00[a]	1.00[a]	2.448e-06	η_p^2	25.48
				sr^2_{Group}	20.41

Notes: [a] Adjusted means; sr^2_{Group} = uniqueness coefficient

Table V.
Comparison of simulated Dataset 2 results by analysis type of posttest scores

Analysis	Control *M*	Intervention *M*	*p*-value for group	Effect size type	Effect size (%)
t-test	3.00	4.00	2.596e-05	*d*	1.00
ANOVA	3.00	4.00	2.596e-05	η^2	20.41
ANCOVA	3.00[a]	4.00	2.101e-06	η_p^2	25.48
Regression	3.00[a]	4.00[a]	2.448e-06	η_p^2	25.48
				sr^2_{Group}	20.41

Notes: [a] Adjusted means; sr^2_{Group} = uniqueness coefficient

identified the studies as the following: eight quasi-experiments, four experiments, two factorial designs, two experimental and control experiments, one controlled experiment, one independent sample experiment, one quasi-field experiment, one empirical experiment, one true experiment and one single-case multiple-baseline participants experimental design. Of the 22 studies identified as an experiment, eight collected data from pretest-posttest control designs with two groups and were coded to adhering to recommended reporting practices and reporting of associated statistical assumptions. Table VI includes the coding results. All forthcoming references to studies correspond to the study numbers in the first column of Table VI. Four of the eight studies (3, 4, 7 and 8) reported random group assignment.

Statistical tests, unique assumptions and select descriptive statistics

We report type of tests coded, statistical assumptions unique to each test reported as well as what measures of central tendency and dispersion were reported. Type of tests coded included *t*-tests ($k = 3$), ANCOVA ($k = 1$) and MLR ($k = 1$), in addition to three others. Only Studies 1 and 4 reported *SD*s disaggregated by group and measurement occasion.

T-*tests*. Two studies (Studies 1 and 2) were coded as *t*-tests with gain scores as the dependent variable. Study 1 reported findings from a repeated-measures ANOVA including main effects and the interaction effect. The former are uninteresting for the pretest-posttest control group design and the latter we coded as a *t*-test of gain scores (Huck and McLean, 1975). Raw mean scores for both control and experimental groups were reported; no baseline differences were identified. Study 2 reported findings from a *t*-test with gain scores. Mean gain scores were reported for both control and experimental groups; pretest differences were identified. Homogeneity of regression slopes and homogeneity of variance were not reported in either Study 1 or Study 2.

Study 3 reported findings from a *t*-test with posttest scores as the dependent variable. Raw mean scores for pretest and posttest scores were disaggregated by control and experimental group, and pretest differences were tested and no statistically significant differences were identified. Homogeneity of variance was not reported.

ANCOVA and MLR. Study 4 reported findings from a study with random assignment of participants to group and where posttest scores were analyzed with ANCOVA. No assumptions unique to ANCOVA were reported. Raw mean scores for pretest and posttest data were reported and disaggregated by control and experimental group. Study 5 reported findings from an MLR with posttest as the dependent variable and group, moderator and interaction between moderator and group as independent variables. Homoscedasticity of residuals and linearity assumptions were not reported. Raw mean pretest and posttest scores for the moderator variable were reported, disaggregated by experimental and control group.

Other tests. In Study 6, even though pretest and posttest data were collected for both control and experimental groups, only posttest data for experimental were analyzed with a one-way ANOVA that we coded as a *t*-test for experimental data (*t*-test-exp). Raw mean pretest and posttest scores were reported for the experimental group. Study 7 reported main effects from a repeated-measures ANOVA. Raw posttest mean scores for both groups were reported, but pretest mean scores were aggregated across groups. Study 8 reported findings from separate paired *t*-tests for the control group and

Table VI.
Pretest-posttest group design studies

Study	Statistical test coded	Disaggregated by group and measurement occasion		Reliability coefficient	Independence of observations	Normality	Random group assignment	Unique statistical test assumptions[b]	Group difference effect size
		M	SD						
1	*t*-test	✓	✓	×	×	×	×	×	×
2	*t*-test	✓	×	×	×	×	×	×	×
3	*t*-test	✓	×	×	×	×	✓	PE	×
4	ANCOVA	✓[c]	✓	×	×	×	✓	×	✓
5	MLR	✓	×	×	×	×	×	×	×
6	*t*-test-exp[a]	✓	×	×	×	×	×	×	×
7	RM ANOVA main effects	×	×	×	×		✓		NA
8	Paired *t*-test	×	×	×	✓		✓		NA

Notes: PE = pretest equivalence only; [a]experimental group only; [b]not coded for RM ANOVA or paired *t*-tests; [c]adjusted mean scores not reported

experimental group and then compared the statistical significance results. Raw means were reported for pretest and posttest scores but were aggregated across groups.

Normality, reliability and independence assumptions

Independence of observations, normality and reliability are assumptions common of all of the analyses listed in Table II. We therefore coded each of the eight tests to determine if researchers addressed these assumptions. The normality assumption was not addressed in any of the eight studies.

Reliability. In our review of the eight studies, researchers tended to report reliability coefficients but did not necessarily report reliability coefficients for data disaggregated by measurement occasion (i.e. pretest, posttest) or group (i.e. control, experimental). In Study 6 that limited analyses to pretest-posttest differences for the experimental group, ranges of reliability coefficients were reported. In Study 5, reliability coefficients were reported for each construct aggregated across groups, but the researchers did not report whether the reliability coefficients were for pretest, posttest or across both measurement occasions. In Studies 3, 4 and 8, researchers reported reliabilities for both pretest and posttest, but coefficients were aggregated across groups. Study 1 reported range of alphas for constructs but did not specify if alpha was for pretest or posttest scores. Studies 2 and 7 did not report any reliability coefficients; however, Study 7 analyzed data from a performance test.

Independence of observations. Independence of observations was discussed in one of the eight studies (Study 8); in that study, researchers purposely selected participants to ensure independence. However, violations of this assumption did not appear evident in five studies, was unclear in one study and was evident in one study. In the latter case, no analysis (e.g. hierarchical linear modeling) that considered the clustered nature of the data appeared to be conducted.

Effect sizes and confidence intervals

For the statistical analyses examining differences between groups, effect sizes were not reported in five studies, but confidence intervals were reported in Study 2. Of the other three studies, Study 4 reported an effect size for the analysis examining differences between groups and two reported effect sizes for main test for a repeated-measures ANOVA but not for the interaction F (i.e. Studies 1 and 7).

Discussion

Journals are an important source of information for researchers and make a lasting impact because results are archived for future reference. Findings from HRD research can have practical implications on the workplace and employees' futures, particularly when human resource departments make decisions based on published findings. Therefore, the HRD community has the responsibility of protecting the integrity of HRD research. Although "pretest-posttest control group design (or an extension of it) is a highly prestigious experimental design", the analysis of pretest-posttest data can be confusing (Huck and McLean, 1975, p. 511).

Recommendations

We collected data from pretest-posttest designs with two groups from select HRD publications. Based on those findings and our illustrative examples, we offer three recommendations relative to pretests-posttest control group designs. First, authors should:

report effect sizes and confidence intervals, report associated statistical assumptions and use appropriate analyses when assumptions fail, and provide descriptive statistics including reliability coefficients for each measurement occasion and group. Second, researchers should become informed consumers of repeated-measures ANOVA. Third, researchers should consider MLR as a statistical analysis to isolate intervention effects.

Effect sizes and confidence intervals. Practical significance needs to be considered through the reporting of effect sizes and confidence intervals (American Psychological Association, 2001; Thompson, 2006a, 2006b). This is because statistical significance does not often provide the answer to the question most researchers wish to answer. As noted by Kirk (1996, p. 746):

> […] statistical significance is concerned with whether a research result is due to chance or sampling variable; practical significance is concerned with whether the result is useful in the real world.

Confidence intervals need to be reported because they provide more information than null hypothesis statistical testing and encourage meta-analytic thinking (Zientek *et al.*, 2010; Thompson, 2006a). As noted by Thompson (2007, p. 427), "CIs are extremely useful, because they convey not only our point estimate, but also, via the width of the intervals, something about the precision of our estimates".

Statistical assumptions. Throughout the analysis stage, assumptions need to be checked and reported (Williams *et al.*, 2013). Common assumptions include independence of observations, normality and reliability. Appropriate homogeneity assumptions should be considered for each test, which include homogeneity of variance for *t*-tests, homoscedasticity of residuals for ANCOVA and MLR and homogeneity of regression slopes for ANCOVA. In the articles we reviewed, HRD researchers were not reporting many of the associated statistical assumptions (Table II). Our illustrations demonstrate the importance of testing assumptions and the need for researchers to consider alternate analyses when assumptions fail, particularly the homogeneity of regression slopes assumption. As well, it is important for researchers to report how data meet the assumption of independence of observations and disaggregated descriptive statistics.

Homogeneity of regression slopes. For ANCOVAs, homogeneity of regression slopes is important because the scores on the control group's and intervention group's dependent variable are calculated with a pooled regression slope and "if the groups' individual slopes differ sharply, then the pooling becomes a muddy average" (Owen and Froman, 1998, p. 559). Homogeneity of regression slopes is just as important for gain scores. Note that:

> […] a gain score analysis is identical to an analysis of covariance except that an a priori decision is made to set b_w to 1.00 rather than let the data dictate the value of this constant (Huck and McLean, 1975, p. 517).

Ignoring the homogeneity of regression slopes assumptions can lead to "tragically misleading analyses" (Campbell and Erlebacher, 1975, p. 597).

Independence of observations. Another assumption that needs to be reported is independence of observations. Complex data such as employees within teams and then teams within departments or employees from departments where departments are organized into units do not meet the independence of observation assumption. Employee opinions in a given unit might be similar and, therefore, perceptions of

employees in different departments within the same unit cannot be assumed to be independent. As noted by Nimon (2011), when the independence of observation assumption fails, then "a statistical test that models the nonindependence (e.g. multilevel modeling)" should be conducted or evidence should be provided "that the structure of the data (i.e. employees nested within departments and departments within organizations) does not impact the accuracy of the statistic reported" (p. 388). As noted previously, failing to consider the nested structure of data can lead to misestimates and even the reporting of opposite findings (Lane *et al.*, 2012; Osborne, 2000).

Descriptive statistics. Means, standard deviations and reliability coefficients should be reported for both pretest and posttest data and should be disaggregated by groups, which will allow readers to make their own informed decisions regarding differences between group and measurement occasions. Because dispersion statistics (e.g. *SD*s) indicate how well measures of central tendency represent the data, researchers should "*always* report the *SD* whenever reporting the mean" (Thompson, 2006a, p. 72). In addition, a full complement of descriptive statistics, including correlation matrices, should be reported, as such statistics might encourage meta-analytic thinking and allow for secondary analyses, as was conducted in the present article (Chermack *et al.*, 2015; Zientek and Thompson, 2009). Although it was encouraging that most of the HRD researchers appeared to understand the importance of reporting reliability coefficients (Nimon, 2012; Nimon *et al.*, 2012; Thompson, 2003), it is important to remember that score reliability is a property of data and should be calculated for each group and measurement occasion (Onwuegbuzie *et al.*, 2005). Researchers reporting on the results of ANCOVA or a residualized dependent variable, which is not the same nomologically as the observed dependent variable, should also report adjusted group means and reference the dependent variable as an adjusted or residualized variable (Nimon and Henson, 2015; Tracz *et al.*, 2005).

Repeated-measures ANOVA

Our findings indicate that 40 years since the publication of Huck and McLean (1975), some researchers did not focus on the interaction result of a repeated-measures ANOVA when analyzing pretest-posttest designs with two groups. Repeated-measures ANOVA is confusing because three *F*s ratios are reported, but the interaction *F* is the only one potentially worth reporting. As noted by Huck and McLean (1975), the *F* for the main effect of the between-subjects factor is of "little utility since it underestimates the variability of the treatment effects" and the *F* from the repeated-measures ANOVA is "worthless from an experimental point of view" (p. 515). Only the interaction *F* is potentially useful; however, "it will always be equal to the gain score *F*" (Huck and McLean, 1975, p. 515).

Multiple linear regression

Our results provide support for examining pretest-posttest control group data with MLR. First, MLR subsumes *t*-test and ANCOVA (Zientek and Thompson, 2009). Second, MLR can be used when the homogeneity of regression slope assumption fails by including an interaction term between group and pretest (Gliner *et al.*, 2003; Linn and Slinde, 1977). In either case, the results of MLR can be used to compute a η_p^2 or uniqueness coefficient to indicate the effect associated with the intervention. The η_p^2 indicates how much of the residualized dependent variable (i.e. dependent variable

where all variance associated with the pretest has been eliminated) is associated with the intervention, whereas the uniqueness coefficient indicates how much of the original dependent variable is associated with the intervention. The choice of which to report is up to the researcher and his or her concern regarding the nomological validity of residualized dependent variable (Nimon and Henson, 2015; Thompson, 2006a; Tracz *et al.*, 2005).

Limitations and suggestions for future research

As the social and behavioral science field continues to evolve, researchers are introducing new methods related to pretest-posttest designs. One limitation of our review is that we did not cover the Johnson–Neyman analysis, which is an alternative to ANCOVA when homogeneity of regression slopes fails (D'Alonzo, 2004). Another limitation is that we did not provide extensions to include follow-up designs (Mara *et al.*, 2012; McArdle, 2009; Mun *et al.*, 2009; Willoughby *et al.*, 2007). In addition, we did not address multivariate analyses, structural equation modeling or hierarchical linear modeling. We also limited our analyses of articles to those where authors explicitly identified the design of their studies as experiment. Further, we limited the discussion to two groups with data considered to be interval (Norman, 2010). Possible future research might include exploring areas we did not consider as well as investigating bias of effect sizes when assumptions fail for *t*-tests and ANCOVAs.

Conclusion

Researchers aim to produce study results that will be replicable. If an incorrect analysis is conducted because assumptions are not met, then the intervention effect might be overestimated or underestimated; thus, results might not be replicable. The HRD field is obligated to publishing research that is accurate and warranted. The scenario we provided in the illustration presented in Table III reminds us of the nursery rhyme character Goldilocks. First, one effect size was too big, then one effect size was too small, then one effect size was just right. However, consequences of trying out different analyses has implications for the field that extend far beyond the inconveniences experienced by Goldilocks and her stakeholders.

Note

1. As noted by Henson (2006), Cohen's *d*, a standardized mean difference, can be converted into a variance-accounted-for effect size using Aaron *et al.*'s (1998) formula: $r = d/\sqrt{d^2 + N^2 - 2N/n1n2}$, where *N* is the total sample size and *n1* and *n2* are the sample size for the two groups, respectively.

References

Aaron, B., Kromrey, J.D. and Ferron, J. (1998), "Equating 'r'-based and 'd'-based effect size indices: problems with a commonly recommended formula", paper presented at the annual meeting of the Florida Educational Research Association, Orlando.

American Psychological Association (2001), *Publication Manual of the American Psychological Association*, 5th ed., Washington, DC.

Boneau, C.A. (1960), "The effects of violations of assumptions underlying the *t* test", *Psychological Bulletin*, Vol. 37, pp. 49-64, doi: 10.1037/h0041412.

Campbell, D.T. and Erlebacher, A. (1975), "How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful", in Guttentag, M. and Struening, E.L. (Eds), *Handbook of Evaluation Research*, Sage, Beverly Hills, CA, Vol. 1, pp. 597-617.

Chermack, T., Coons, L.M., Nimon, K., Bradley, P. and Glick, M.B. (2015), "The effects of scenario planning on participant perceptions of creative organizational climate", *Journal of Leadership & Organizational Studies*, Vol. 22, pp. 355-371, doi: 10.1177/1548051815582225.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Lawrence Erlbaum Associates, Hillsdale, NJ.

Cronbach, L.J. and Furby, L. (1970), "How should we measure 'change': or should we?", *Psychological Bulletin*, Vol. 74, pp. 68-80, doi: 10.1037/h0029382.

D'Alonzo, K.T. (2004), "The Johnson-Neyman procedure as an alternative to ANCOVA", *Western Journal of Nursing Research*, Vol. 26, pp. 804-812, doi: 10.1177/0193945904266733.

Edwards, A.L. (1960), *Experimental Design in Psychological Research*, Holt, Rinehart & Winston, New York, NY.

Forbes, A.B. and Carlin, J.B. (2005), "Residual change analysis is not equivalent to analysis of covariance", *Journal of Clinical Epidemiology*, Vol. 58, pp. 540-541, doi: 10.1016/j.jclinepi.2004.12.002.

Gliner, J.A., Morgan, G.A. and Harmon, R.J. (2003), "Pretest-posttest comparison group designs: analysis and interpretation", *Journal of the American Academy of Child & Adolescent Psychiatry*, Vol. 42, pp. 500-503, doi: 0.1097/01.CHI.0000046809.95464.BE.

Hamilton, B.L. (1977), "An empirical investigation of the effects of heterogeneous regression slopes in analysis of covariance", *Educational and Psychological Measurement*, Vol. 37, pp. 701-712, doi: 10.1177/001316447703700313.

Harwell, M. (2003), "Summarizing Monte Carlo results in methodological research: the single-factor, fixed-effects ANCOVA case", *Journal of Educational and Behavioral Statistics*, Vol. 28, pp. 45-70, doi: 10.3102/10769986028001045.

Henson, R.K. (1998), "ANCOVA with intact groups: don't do it!", paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, LA.

Henson, R.K. (2006), "Effect-size measures and meta-analytic thinking in counseling psychology research", *The Counseling Psychologist*, Vol. 34, pp. 601-629, doi: 10.1177/0011000005283558.

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

Humphreys, L.G. (1976), "Analysis of data from pre-and posttest designs. A comment", *Psychological Reports*, Vol. 38, pp. 639-642, doi: 10.2466/pr0.1976.38.2.639.

Isaksen, S.G. and Akkermans, H.J. (2011), "Creative climate: a leadership lever for innovation", *Journal of Creative Behavior*, Vol. 45, pp. 161-187, doi: 10.1002/j.2162-6057.2011.tb01425.x.

Jamieson, J. (2004), "Analysis of covariance (ANCOVA) with difference scores", *International Journal of Psychophysiology*, Vol. 52, pp. 277-283, doi: 10.1016/j.ijpsycho.2003.12.009.

Kirk, R.E. (1996), "Practical significance: a concept whose time has come", *Educational and Psychological Measurement*, Vol. 56, pp. 746-759, doi: 10.1177/0013164496056005002.

Kisbu-Sakarya, Y., MacKinnon, D.P. and Aiken, L.S. (2013), "A Monte Carlo comparison study of the power of the analysis of covariance, simple difference, and residual change scores in testing two-wave data", *Educational and Psychological Measurement*, Vol. 73, pp. 47-62, doi: 10.1177/0013164412450574.

Knapp, T.R. and Schafer, W.D. (2009), "From gain score t to ANCOVA F (and vice versa)", *Practical Assessment, Research and Evaluation*, Vol. 14, pp. 1-8, available at: http://pareonline.net.ezproxy.shsu.edu/getvn.asp?v=14&n=6

Kraha, K., Turner, H., Nimon, K., Zientek, L.R. and Henson, R. (2012), "Tools to support interpreting multiple regression in the faculty of multicollinearity", *Frontiers in Measurement*, Vol. 3, p. 44, doi: 10.3389/fpsyg.,2012.00044.

Lane, F., Nimon, K. and Roberts, J.K. (2012), "A random intercepts model of part – time employment and standardized test scores using SPSS", in Garson, G.D. (Ed.), *Hierarchical Linear Modeling: Guide and Applications*, Sage, Thousand Oaks, CA, pp. 149-165.

Levy, K.J. (1980), "A Monte Carlo study of analysis of covariance under violations of the assumptions of normality and equal regression slopes", *Educational and Psychological Measurement*, Vol. 40, pp. 835-840, doi: 10.1177/001316448004000404.

Linn, R.L. and Slinde, J.A. (1977), "The determination of the significance of change between pre-and posttesting periods", *Review of Educational Research*, Vol. 47, pp. 121-150, doi: 10.3102/00346543047001121.

Lord, F.M. (1963), "Elementary models for measuring change", in Harris, C.W. (Ed.), *Problems in Measuring Change*, University of Wisconsin Press, Madison, pp. 21-38.

McArdle, J.J. (2009), "Latent variable modeling of differences and changes with longitudinal data", *Annual Review of Psychology*, Vol. 60, pp. 577-605, doi: 10.1146/annurev.psych.60.110707.163612.

Mara, C.A., Cribbie, R.A., Flora, D.B., LaBrish, C., Mills, L. and Fiksenbaum, L. (2012), "An improved model for evaluating change in randomized pretest, posttest, follow-up designs", *Methodology*, Vol. 8 No. 3, pp. 97-103, doi: 10.1027/1614-2241/a000041.

Maxwell, S.E. and Delany, H.D. (2004), *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, 2nd ed., Lawrence Erlbaum, Mahwah, NJ.

Maxwell, S.E., Delaney, H.D. and Manheimer, J.M. (1985), "ANOVA of residuals and ANCOVA: correcting an illusion by using model comparisons and graphs", *Journal of Educational Statistics*, Vol. 10, pp. 197-200, doi: 10.2307/1164792.

Miller, G.A. and Chapman, J.P. (2001), "Misunderstanding analysis of covariance", *Journal of Abnormal Psychology*, Vol. 110, pp. 40-48, doi: 10.1037/0021-843X.110.1.40.

Mun, E.Y., von Eye, A. and White, H.R. (2009), "An SEM approach for the evaluation of intervention effects using prepost-post designs", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 16, pp. 315-337, doi: 10.1080/10705510902751358.

Nimon, K. (2011), "Improving the quality of quantitative research reports: a call for action", *Human Resource Development Quarterly*, Vol. 22, pp. 387-394, doi: 10.1002/hrdq.20091.

Nimon, K. (2012), "Statistical assumptions of substantive analyses across the general linear model: a mini-review", *Frontiers in Quantitative Psychology and Measurement*, Vol. 3, p. 322, doi: 10.3389/fpsyg.2012.00322.

Nimon, K. and Henson, R.K. (2015), "Validity of a residualized dependent variable after pretest covariance adjustments: still the same variable?", *Journal of Experimental Education*, Vol. 83, pp. 405-422, doi: 10.1080/00220973.2014.907228.

Nimon, K., Zientek, L.R. and Henson, R.K. (2012), "The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data", *Frontiers in Quantitative Psychology and Measurement*, Vol. 3, p. 102, doi: 10.3389/fpsyg. 2012.00102.

Norman, G. (2010), "Likert scales, levels of measurement and the 'laws' of statistics", *Advances in Health Sciences Education*, Vol. 15, pp. 625-632, doi: 10.1007/s10459-010-9222-y.

Onwuegbuzie, A.J., Roberts, J.K. and Daniel, L.G. (2005), "A proposed new 'what if' reliability analysis for assessing the statistical significance of bivariate relationships", *Measurement and Evaluation in Counseling and Development*, Vol. 37, pp. 228-239.

Osborne, J. (2000), "Advantages of hierarchical linear modeling", *Practical Assessment, Research & Evaluation*, Vol. 7 No. 1, available at: http://pareonline.net/getvn.asp?v_7&n_1

Overall, J.E. and Woodward, J.A. (1975), "Unreliability of difference scores: a paradox for measurement of change", *Psychological Bulletin*, Vol. 82, pp. 85-86, doi: 10.1037/h0076158.

Owen, S.V. and Froman, R.D. (1998), "Uses and abuses of the analysis of covariance", *Research in Nursing & Health*, Vol. 21, pp. 557-562, doi: 10.1002/(SICI)1098-240X(199812)21:6%3C557::AID-NUR9%3E3.0.CO;2-Z.

Pedhazur, E.J. (1997), *Multiple Regression in Behavioral Research: Explanation and Prediction*, 3rd ed., Harcourt Brace, Fort Worth, TX.

Rogosa, D., Brandt, D. and Zimowski, M. (1982), "A growth curve approach to the measurement of change", *Psychological Bulletin*, Vol. 92, pp. 726-748, doi: 10.1037/0033-2909.92.3.726.

Russ-Eft, D. and Hoover, A.L. (2005), "Experimental and quasi-experimental designs", in Swanson, R.A. and Holton, E.F. (Eds), *Research in Organizations: Foundation and Methods in Inquiry*, Berrett-Koehler, San Francisco, CA, pp. 75-96.

Thompson, B. (2006a), *Foundations of Behavioral Statistics: An Insight-Based Approach*, Guilford, New York, NY.

Thompson, B. (2006b), "Research synthesis: effect sizes", in Green, J.L., Camilli, G. and Elmore, P.B. (Eds), *Handbook of Complementary Methods in Education Research*, Erlbaum, Mahwah, NJ, pp. 583-603.

Thompson, B. (2007), "Effect sizes, confidence intervals, and confidence intervals for effect sizes", *Psychology in the Schools*, Vol. 44, pp. 423-432.

Thompson, B. (Ed.) (2003), *Score Reliability: Contemporary Thinking on Reliability Issues*, Sage, Newbury Park, CA.

Tracz, S.M., Nelson, L.L., Newman, I. and 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*, Vol. 31, pp. 16-21.

Williams, M.N., Grajales, C.A.G. and Kurkiewicz, D. (2013), "Assumptions of multiple regression: correcting two misconceptions", *Practical Assessment, Research & Evaluation*, Vol. 18, pp. 1-14, available at: http://pareonline.net/getvn.asp?v=18&n=11

Williams, R.H. and Zimmerman, D.W. (1996), "Are simple gain scores obsolete?", *Applied Psychological Measurement*, Vol. 20, pp. 59-69, doi: 10.1177/014662169602000106.

Willoughby, M., Vandergrift, N., Blair, C. and Granger, D.A. (2007), "A structural equation modeling approach for the analysis of cortisol data collected using pre-post-post designs", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 14, pp. 125-145, doi: 10.1080/10705510709336740.

Wu, Y.-W.B. (1984), "The effects of heterogeneous regression slopes on the robustness of two test statistics in the analysis of covariance", *Educational and Psychological Measurement*, Vol. 44, pp. 647-663, doi: 10.1177/0013164484443011.

Zientek, L.R. and Thompson, B. (2009), "Matrix summaries improve research reports: secondary analyses using published literature", *Educational Researcher*, Vol. 38, pp. 343-352, doi: 10.3102/0013189X09339056.

Zientek, L.R., Yetkiner, Z.E. and Thompson, B. (2010), "Characterizing the mathematics anxiety literature using confidence intervals as a literature review mechanism", *The Journal of Educational Research*, Vol. 103, pp. 424-438, doi: 10.1080/00220670903383093.

Zimmerman, D.W. and Williams, R.H. (1982), "Gain scores in research can be highly reliable", *Journal of Educational Measurement*, Vol. 19, pp. 149-154, doi: 10.1111/j.1745-3984.1982.tb00124.x.

Zumbo, B.D. (1999), "The simple difference score as an inherently poor measure of change: some reality, much mythology", in Thompson, B. (Ed.), *Advances in Social Science Methodology*, JAI Press, Greenwich, CT, Vol. 5, pp. 269-304.

Appendix 1. R code to replicate analyses for Dataset 1

```
###Load necessary packages
library(yhat)
library(car)
library(lattice)
library(effects)
library(effsize)
library(MASS)
library(gdata)

###Set contrasts
options(contrasts=c("contr.sum",'contr.poly'))

###Create simulated dataset from descriptive statistics reported in
###Chermack et al. (2015)

###Experimental Simulated Data
expcov<-matrix(c(0.3643494, 0.1556344, 0.1556344, 0.3928290),2,2)
rownames(expcov)<-colnames(expcov)<-c("PlayHumorPre","PlayHumorPost")
expdata<-mvrnorm(n=48,c(3.02,3.06),expcov,empirical=TRUE)
expdata<-data.frame(expdata)
expdata$Group<-1

###Control Simulated Data
ctlcov<-matrix(c(0.26082398, -0.09575429, -0.09575429,0.43044264),2,2)
rownames(ctlcov)<-colnames(ctlcov)<-c("PlayHumorPre","PlayHumorPost")
ctldata<-mvrnorm(n=42,c(2.55,2.56),ctlcov,empirical=TRUE)
ctldata<-data.frame(ctldata)
ctldata$Group<-0

###Merged Simulated Data
Dataset<-rbind(expdata,ctldata)
Dataset$Group<-as.factor(Dataset$Group)
levels(Dataset$Group)<-c("Control","Treatment")
Dataset$PlayHumorPrec<-Dataset$PlayHumorPre-mean(Dataset$PlayHumorPre,na.rm=TRUE)

###Run descriptive statistics
aggregate(PlayHumorPre~Group,Dataset,mean)
aggregate(PlayHumorPre~Group,Dataset,sd)
aggregate(PlayHumorPost~Group,Dataset,mean)
aggregate(PlayHumorPost~Group,Dataset,sd)

###Check pretest differences
(aoutpr<-anova(aov(PlayHumorPre~Group,Dataset)))
aoutpr[1,"Sum Sq"]/sum(aoutpr[,"Sum Sq"])

###Check homogeneity of regression slope statistically
aouth<-aov(PlayHumorPost~Group*PlayHumorPre,Dataset)
```

```
summary(aouth)
###Check homogeneity of regression slope visually

xyplot(PlayHumorPre~PlayHumorPost|Group,layout=c(2,1),col="black",
   type=c("p", "r"),data=Dataset)

###Check homogeneity of regression slope descriptively
sDataset<-split(Dataset,Dataset$Group)
cor(sDataset$Control$PlayHumorPre,sDataset$Control$PlayHumorPost)
cor(sDataset$Treatment$PlayHumorPre,sDataset$Treatment$PlayHumorPost)

###Use regression to analyze
lmout<-lm(PlayHumorPost~Group*PlayHumorPrec,data=Dataset)
Anova(lmout,type=3)
effect("Group",lmout)
regr(lmout)

###Use t-test to analyze
t.test(PlayHumorPost~Group,var.equal=TRUE,data=Dataset)
cohen.d(PlayHumorPost~Group,pooled=TRUE,data=Dataset)

###Use ANOVA to analyze
(aoutpo<-anova(aov(PlayHumorPost~Group,Dataset)))
aoutpo[1,"Sum Sq"]/sum(aoutpo[,"Sum Sq"])
effect("Group",aov(PlayHumorPost~Group,Dataset))

###Use ANCOVA to analyze
(ancout<-Anova(lm(PlayHumorPost~PlayHumorPre+Group,Dataset),type="III"))
ancout["Group","Sum Sq"]/(ancout["Group","Sum Sq"]+ancout["Residuals","Sum Sq"])
effect("Group",lm(PlayHumorPost~PlayHumorPre+Group,Dataset))
```

Appendix 2. R code to replicate analyses for Dataset 2

```
###Load necessary packages
library(foreign, pos=4)
library(yhat)
library(car)
library(lattice)
library(effects)
library(effsize)
library(MASS)
library(gdata)

###Set contrasts
options(contrasts=c("contr.sum",'contr.poly'))

###Create simulated dataset
###Experimental Simulated Data

expcov<-matrix(c(1, .5, .5, 1),2,2)
rownames(expcov)<-colnames(expcov)<-c("Pre","Post")
expdata<-mvrnorm(n=40,c(3.00,4.00),expcov,empirical=TRUE)
expdata<-data.frame(expdata)
expdata$Group<-1

###Control Simulated Data
```

```
ctlcov<−matrix(c(1, .5, .5, 1),2,2)
rownames(ctlcov)<−colnames(ctlcov)<−c("Pre","Post")
ctldata<−mvrnorm(n=40,c(3.00,3.00),ctlcov,empirical=TRUE)
ctldata<−data.frame(ctldata)
ctldata$Group<−0

###Merged Simulated Data
Dataset<−rbind(expdata,ctldata)
Dataset$Group<−as.factor(Dataset$Group)
levels(Dataset$Group)<−c("Control","Treatment")
Dataset$Prec<−Dataset$Pre-mean(Dataset$Pre,na.rm=TRUE)
Dataset$Diff<−dataset$Post-Dataset$Pre

###Run descriptive statistics
aggregate(Pre~Group,Dataset,mean)
aggregate(Pre~Group,Dataset,sd)
aggregate(Post~Group,Dataset,mean)
aggregate(Post~Group,Dataset,sd)
aggregate(Diff~Group,Dataset,mean)
aggregate(Diff~Group,Dataset,sd)

###Check pretest differences
(aoutpr<−anova(aov(Pre~Group,Dataset)))
aoutpr[1,"Sum Sq"]/sum(aoutpr[,"Sum Sq"])

###Check homogeneity of regression slope statistically
aouth<−aov(Diff~Group*Pre,Dataset)
summary(aouth)

###Check homogeneity of regression slope visually
xyplot(Pre~Diff|Group,layout=c(2,1),col="black",type=c("p",
"r"),xlab="Difference",ylab="Pretest",data=Dataset)

###Check homogeneity of regression slope descriptively
sDataset<−split(Dataset,Dataset$Group)
cor(sDataset$Control$Pre,sDataset$Control$Diff)
cor(sDataset$Treatment$Pre,sDataset$Treatment$Diff)

###Analysis of difference scores

###Use regression to analyze
lmout<−lm(Diff~Group*Prec,data=Dataset)
Anova(lmout,type=3)
effect("Group",lmout)
regr(lmout)

###Use t-test to analyze
t.test(Diff~Group,var.equal=TRUE,data=Dataset)
cohen.d(Diff~Group,pooled=TRUE,data=Dataset)

###Use ANOVA to analyze
(aoutpo<−anova(aov(Diff~Group,Dataset)))
aoutpo[1,"Sum Sq"]/sum(aoutpo[,"Sum Sq"])
effect("Group",aov(Diff~Group,Dataset))

###Use ANCOVA to analyze
```

```
(ancout<−Anova(lm(Diff~Pre+Group,Dataset),type="III"))
ancout["Group","Sum Sq"]/(ancout["Group","Sum Sq"]+ancout["Residuals","Sum Sq"])
effect("Group",lm(Diff~Pre+Group,Dataset))

###Analysis of posttest scores

###Use regression to analyze
lmout<−lm(Post~Group*Prec,data=Dataset)
Anova(lmout,type=3)
effect("Group",lmout)
regr(lmout)

###Use t-test to analyze
t.test(Post~Group,var.equal=TRUE,data=Dataset)
cohen.d(Post~Group,pooled=TRUE,data=Dataset)

###Use ANOVA to analyze
(aoutpo<−anova(aov(Post~Group,Dataset)))
aoutpo[1,"Sum Sq"]/sum(aoutpo[,"Sum Sq"])
effect("Group",aov(Post~Group,Dataset))

###Use ANCOVA to analyze
(ancout<−Anova(lm(Post~Pre+Group,Dataset),type="III"))
ancout["Group","Sum Sq"]/(ancout["Group","Sum Sq"]+ancout["Residuals","Sum Sq"])
effect("Group",lm(Post~Pre+Group,Dataset))
```

Corresponding author

Linda Zientek can be contacted at: lrzientek@shsu.edu