A Comparison of Difference Score and Pre-test Post-test Repeated Measures ANOVA: Implications for Research in Management Information Systems

Dr. Thomas A. Wright, Department of Managerial Sciences, University of Nevada

Abstract

Various forms of the pretest-posttest design are extensively used in Management Information Systems (MIS) research. There is a widespread misconception among MIS researchers regarding the equivalence of the difference score and pretest-posttest repeated measures ANOVA. Several important implications of the equivalence are presented which concern the interpretation of interaction effects, the test of normality, the test of equality of variance, and the experimentwise Type I error rate.

Introduction

Writing over 20 years ago, Vroom (1966) identified potential benefits of various "dynamic" versus "static" methods of analysis in the study of organizational phenomena. Vroom defined as dynamic, those designs with measurements at two points in time. Consider the following study as an example. Two groups of end users are defined relative to whether they successfully completed a training course in hardware use. Measurement of their job performance is obtained at two time periods, identified as pretraining (T_1) and posttraining (T_2), separated by six months. Pretest-posttest designs of this type are commonplace in the management information systems (MIS) literature (Baronas & Louis, 1988; Kaplan & Duchon, 1988; Headrick & Morgan, 1989). In this example, the MIS researcher wishes to test the null hypothesis that performance change over time is the same for both training and no training groups of end users. The null and alternative hypotheses may be stated as

H_o: $\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = 0$
H_1: $\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} \neq 0$

where μ_{ij} is the population mean job performance in Group j at Time i. A test of H_o against H_1 can be obtained using a repeated measures ANOVA with training as the between group factor and Time as the within subject factor (Winer, 1971, section 7.2). The null hypothesis states that the Group by Time interaction effect is equal to zero. Alternatively, a test of H_o against H_1 also can be obtained using a single factor ANOVA with training as the between group factor and the $T_2 - T_1$ difference scores as the dependent variable. The F statistic will be identical in both cases (Brogan & Kutner, 1980).

There are several very important advantages of the difference score ANOVA over the repeated measures ANOVA. However, the algebraic equivalence between these two approaches is not well known among applied business researchers. In a recent review of an article which used the difference score ANOVA, one reviewer for a prestigious applied business journal wrote:

"The analysis might be redone as a 3x2 between/within ANOVA, where the between factor is group membership and the within factor is the repeated T1-T2 measure. Your hypothesis would suggest that you find a Group x Time interaction effect."

In response to a revision in which the authors pointed out the algebraic equivalence of the two approaches, the same reviewer replied:

"The authors assert that using a one-way ANOVA on difference scores is equivalent to examining the Group

x Time interaction. This is an assertion for which I would like to see a reference. I checked with two statisticians regarding the equivalence between these analyses and both said they were not interchangeable."

Clearly, this reviewer as well as the two statisticians are unaware of the equivalence of the difference score and pretest-posttest repeated measures ANOVA. To determine the extent to which this misconception exists among applied management researchers, a survey was conducted in which the above study was described and the respondents were asked if the ANOVA on difference scores or the repeated measures ANOVA would be preferred. Ten professors from eight prestigious Business schools were contacted by telephone. Since it would have been difficult to obtain a truly representative sample from the population of applied management researchers, a sample of applied management researchers with very strong quantitative skills was obtained. Specifically, researcher reputation, publication record, and graduate level training in statistics were used to form the quantitative skill criterion. It was assumed that this sample would over-estimate the proportion of applied management researchers who are aware of the equivalence of the two approaches.

The results of this nonscientific survey are striking. All ten researchers stated that the repeated measures ANOVA would be preferred. Some went on to say that the repeated measures ANOVA is more powerful. One respondent stated that difference scores should never be used.

It is not simply an abstract statistical issue that every researcher in our study and, we presume, a large number of MIS researchers in general, are unaware of the equivalence of the difference score and repeated measures ANOVA. There are four major research implications: 1) the ability of researchers to accurately interpret interaction effects, 2) testing the normality assumption; 3) testing the equality of variance assumption, and 4) experimentwise Type I error rate. These implications are now discussed in detail.

Interpreting Interaction Effects

The difference score and repeated measures ANOVA differ in the order of the interaction effect that must be interpreted. With only one between group factor, the repeated measures ANOVA requires interpretation of a two-way interaction effect. For example, the Training by Time interaction would require interpretation in the illustration given above. In contrast, the difference score ANOVA requires interpretation of the main effect of the between group factor (the difference between training and no training in the above example). Furthermore, with two between group factors, the repeated measures ANOVA requires the interpretation of a three-way interaction while the difference score ANOVA requires interpretation of a two-way interaction. For example, if we consider a pretest-posttest design with Sex of the trainees (male or female) and hardware use Training (absent or present) as the two between group factors, application of the repeated measures ANOVA would require interpretation of a three-way interaction (Sex by Training by Time), while the difference score ANOVA would require interpretation of a two-way interaction (Sex by Training).

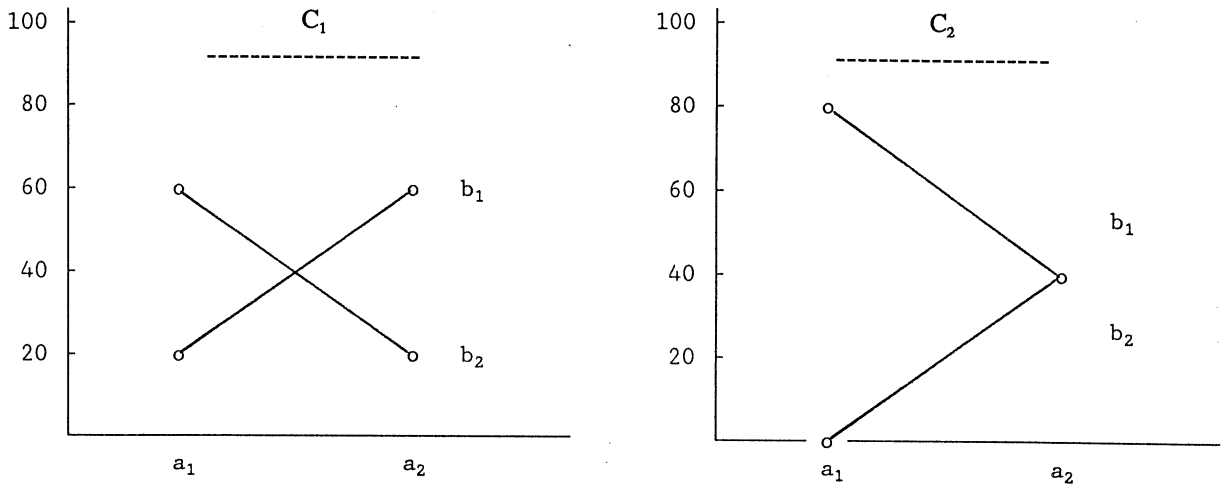
A second survey was conducted in which ten applied management researchers and ten doctoral students majoring in quantitative methods were asked to state the null hypothesis of a zero three-way interaction effect in terms of the population cell means for a balanced 2x2x2 factorial design. None of the respondents gave the correct answer. When asked if they could recognize the presence of a three-way interaction from a line plot of the data, 15 of the 20 respondents answered affirmatively. These 15 respondents were then shown Figure 1 and asked if the data indicated the presence or absence of a three-way interaction effect. Fourteen of the fifteen indicated that a three-way interaction was present. In fact, the data in Figure 1 show a three-way interaction effect that is exactly equal to zero. The null hypothesis of a zero three-way interaction effect can be stated as

$$H_o: \mu_{111} - \mu_{112} - \mu_{121} + \mu_{122} - \mu_{211} + \mu_{212} + \mu_{221} - \mu_{222} = 0$$

Replacing μ_{ijk} in H_o with the means from Figure 1 yields a value of exactly zero. If a researcher cannot state the null hypothesis of a zero three-way interaction, then that researcher will have difficulty interpreting a three-way interaction effect. The results of our survey indicate that the vast majority of applied management researchers are unable to correctly interpret a three-way interaction effect.

The difference score ANOVA involves interaction effects that are one degree lower than those in the repeated measures ANOVA. The difference score ANOVA can be recommended solely on the basis of the fact that the interpretation of a k-way interaction is substantially more difficult than the interpretation of a

Figure 1



(k-1)-way interaction. As a case in point, for the performance study with Sex and hardware use Training as between group factors, the interpretation of the three-way interaction in the repeated measures ANOVA will be substantially more difficult than the interpretation of the two-way interaction using a difference score ANOVA.

Testing the Normality Assumption

In a comprehensive review of normality tests, D'Agnostino (1986) states that the test based on skewness and kurtosis is among the most powerful and easy to apply. The repeated measures ANOVA assumes that the dependent variable is normally distributed at each time period and within each group. This assumption is necessary for tests of all interaction and main effects. Thus, in our performance example, it would be necessary to assess normality within each group and at T_1 and T_2. In contrast, the normality assumption in a difference score ANOVA is satisfied if the difference scores are normally distributed in each treatment population. It is important to note and apparently it is not well known that if the Time by Group interaction effect is the only effect of interest (which is typically the case), then the normality assumption for the repeated measures ANOVA simplifies to the normality assumption of the difference score ANOVA. Normality within each group and at each time period implies normality of difference scores within each group. However, the converse is not true. That is, if univariate normality has been satisfied, the researcher cannot assume that multivariate normality has been satisfied.

Testing the Equality of Variance Assumption

Let Σ denote the 2x2 covariance matrix of the dependent variable at T_1 and T_2. The repeated measures ANOVA assumes that Σ is equal across all subpopulations defined by the between group factors. Furthermore, a sufficient condition for the F statistic to be F distributed in the repeated measures ANOVA is that Σ is compound symmetric. A compound symmetric matrix is a matrix with equal diagonal elements and equal off-diagonal elements. Thus, the compound symmetric and equality of covariance assumptions imply that the variance of the dependent variable is equal across groups and time periods. This assumption must be satisfied since it is known that the repeated measures ANOVA is not robust to a violation of the compound symmetric assumption (Kirk, 1968, p. 142) and the F test is liberal (rejects with probability greater than α) in this case. Unfortunately, the available test for equality of covariance matrices is based on the assumption of multivariate normality and this test is not robust to a violation of the multivariate normality assumption (Seber, 1984, p. 106). This problem is further complicated by the fact that currently available tests of multivariate normality are not entirely satisfactory (D'Agnostino, 1986; Seber, 1984). One approach to testing multivariate normality involves applying the univariate normal tests at each time period. If univariate normality is rejected, then multivariate normality also must be rejected. However, if univariate normality is accepted, multivariate normality does not follow.

Given that widespread knowledge is not apparent, it is important to note that when interest focuses only on the Group by Time interaction, the assumption of compound symmetry and equality of covariance matrices reduces the assumption of equality of variances on the

difference scores (which is the assumption of the difference score ANOVA). This fact is of major importance because a robust small sample test for the equality of variance is available (i.e., Levene's test; see Keppel, 1971). Thus, unlike the repeated measures ANOVA, the underlying assumptions of the difference score ANOVA can be accurately assessed. In addition, with equal sample sizes, the difference score ANOVA is robust to a violation of the equality of variance assumption (Scheffe, 1959, section 10.4).

Experimentwise Type I Error Rate

When more than one hypothesis is tested from a single sample of data, the probability of falsely rejecting one or more null hypothesis is referred to as the *experimentwise Type I error rate.* The greater the number of hypotheses, the larger the experimentwise Type I error rate. This is important because more hypotheses are typically tested in a repeated measures ANOVA than in a difference score ANOVA. Testing the normality assumption in the pretest-posttest repeated measures ANOVA requires twice as many tests as in the difference score ANOVA. In general, the repeated measures ANOVA requires both a test of compound symmetry and a test of equality of covariance matrices compared to the single test of equal variances in the difference score ANOVA. Finally, in the repeated measures ANOVA, more tests of main and interaction effects are typically tested compared with the difference score ANOVA (e.g., in the repeated measures ANOVA with two between group factors, there are six main and interaction effects compared to the three main and interaction effects in the difference score ANOVA).

A simple, but approximate solution to the multiplicity problem involves making a Bonferonni adjustment to the alpha level used in each test. Specifically, alpha is divided by the number of hypotheses that will be tested (Milliken & Johnson, 1984, p. 33). This procedure provides control over the experimentwise Type I error rate but will reduce (sometimes greatly) the power of the test. Application of the difference score ANOVA using the Bonferonni adjustment method is more powerful than the typical application of the repeated measures ANOVA using a Bonferonni adjustment method since the repeated measures ANOVA usually involves more tests of hypotheses than the difference score ANOVA (some of which may not be interesting).

Example

Consider the following data from a two-group, pretest-posttest design with eight end users per group. The dependent variable is the time in seconds to correctly access data.

Table 1
Hypothetical Data for a Two Group, Pretest-posttest Design

Group	T1	T2	T2-T1
	1.1	1.0	-.1
	2.3	2.4	.1
	2.6	2.5	-.1
1	3.3	3.1	-.2
	18.2	18.3	.1
	12.5	12.1	-.4
	4.2	4.5	.3
	3.6	3.4	-.2
	1.2	1.3	.1
	2.4	2.8	.4
	2.5	2.6	.1
2	3.4	3.3	-.1
	3.8	4.5	.7
	4.1	4.4	.3
	11.9	13.2	1.3
	18.9	19.4	.5

Suppose that two different MIS researchers are asked to analyze these data. The first researcher decides to analyze the data as a repeated measures ANOVA. The analysis begins with a test of normality within each group and time period. The skewness and kurtosis tests applied to each group at each time period indicate that the normality assumption is not justified in all four cases. The critical skewness value for eight observations with $\alpha = .05$ is 1.202 (D'Agnostino, 1986, p. 376). The estimated skewness in each group and time period equal 1.28, 1.31, 1.38, and 1.29, all of which exceed the critical value of 1.202. Since univariate normality has been violated, multivariate normality cannot be assumed and the available tests for the equality of covariance matrices and compound symmetry cannot be applied. The F statistic for the Group by Time interaction was computed and was found to equal 7.49 ($p = .016$). However, believing that the assumptions of the repeated measures ANOVA have been violated and knowing that the test is positively biased, this MIS researcher may conclude that there is no real difference between treatments and may abandon the analysis.

The second MIS researcher is aware of the fact that the test of the Group by Time interaction is algebraically

identical to the test of the main effect of Group using T_2 - T_1 difference scores. This researcher begins the analysis with a test of the model assumptions. First, the skewness and kurtosis tests indicate normality within both treatments. The estimated skewness coefficients of .18 and 1.03 for each group are less than the critical skewness value of 1.202 at $\alpha = .05$ and the estimated kurtosis coefficients of 2.25 and 3.22 for each group are within the critical range of 1.40 to 4.09 at $\alpha = .05$. Second, Levene's robust test of equal variances suggests that the homoscedasticity assumption also has been satisfied ($F = 1.77$; $df = 1, 14$; $p = .205$). The MIS researcher then computes the F statistic for the main effect of Group which equals 7.49 ($p = .016$). Since the assumptions of the difference score ANOVA have been satisfied, the second MIS researcher proceeds to write up the results for publication consideration.

Summary

There is a fundamental relationship between the repeated measures ANOVA and the difference score ANOVA. This relationship is not well known among applied management researchers. This is a potentially serious problem since the pretest-posttest design is so widely used and an understanding of this relationship will direct applied researchers toward a more parsimonious and appropriate analysis of their data.

It should be noted that previous research (Cronbach and Furby, 1971; Johns, 1981) have indicated that difference scores should not be used because of their unreliability. While this is an important consideration for correlational studies that do not involve group comparisons, the unreliability of difference scores is not a problem in the difference score ANOVA. Specifically, measurement error in the dependent variable of an ANOVA model has the effect of only reducing the power of the test. Furthermore, it can be shown that the difference score ANOVA is more powerful than an ANOVA on the posttest score when the pretest-posttest correlation is greater than .5.

It is hoped that future research in MIS will incorporate and build upon the pedagogy presented in this article. Given the widespread use of repeated measure designs, a greater understanding and awareness of the appropriateness and equivalence of various forms of repeated measure designs is warranted. Especially so, when the widespread misconception regarding the equivalence of the difference score and pretest-posttest repeated measures ANOVA is noted.

References

1 Baronas, Ann-Marie K. and Meryl Reis Louis, "Restoring a Sense of Control During Implementation: How User Involvement Leads to System Acceptance," *Management Information Systems Quarterly*, Vol. 12, No. 1, pp. 111-124, 1988.

2 Brogan, Donna R. and Michael H. Kutner, "Comparative Analyses of Pretest-Posttest Designs," *The American Statistician*, Vol. 34, No. 4, pp. 229-232, 1980.

3 Cronbach, Lee J. and Lita Furby, "How Should We Measure "Change" - or Should We?" *Psychological Bulletin*, Vol. 74, No. 1, pp. 74-80, 1970.

4 D'Agnostino, Ralph B. "Tests for the Normal Distribution." In Ralph B. D'Agnostino and Michael A. Stephens (Eds.), *Goodness-of-Fit Techniques* (pp. 367-413). New York: Marcel Dekker, 1986.

5 Headrick, R. Wayne and George W. Morgan. "The Effects of Lecture and Individual Study Instructional Methodologies in an Introductory Information Systems Course: A Comparative Analysis," *The Journal of Computer Information Systems Quarterly*, Vol. 29, No. 3, pp. 35-38, 1989.

6 Johns, Gary. "Difference Score Measures of Organizational Behavior Variables: A Critique. *Organizational Behavior and Human Performance*, Vol. 27, No. 3, pp. 443-463, 1981.

7 Kaplan, Bonnie and Dennis Duchon, "Combining Qualitative and Quantitative Methods in Information Systems Research: A Case Study," *Management Information Systems Quarterly*, Vol. 12, No. 4, pp. 571-586, 1988.

8 Keppel, Geoffrey. *Design and Analysis: A Researcher's Handbook.* Englewood Cliffs, N.J.: Prentice-Hall, 1973.

9 Kirk, Roger E. *Experimental Design: Procedures for the Behavioral Sciences.* Belmont, Ca.: Brooks/Cole.

10 Milliken, George A. and Dallas E. Johnson. *The Analysis of Messy Data. Volume I: Designed Experiments.* New York: Van Nostrand Reinhold, 1984.

11 Scheffé, Henry. *The Analysis of Variance.* New York: Wiley, 1959.

12 Seber, George A. *Multivariate Observations.* New York: Wiley, 1984.

13 Vroom, Victor H. "A Comparison of Static and Dynamic Correlational Methods in the Study of Organizations." *Organizational Behavior and Human Performance*, Vol.1, No. 1, pp. 55-70, 1966.

14 Winer, Bernard J. *Statistical Principles in Experimental Design.* New York, McGraw-Hill, 1971.