On Analyzing Repeated Measures Designs With Both Univariate and Multivariate Methods: A Primer with Examples

Kevin M. Kieffer
Saint Leo University
James A. Haley VA Medical Center, Tampa

The present paper provides an introductory exposure to different approaches currently available for analyzing data generated through repeated measurements of a phenomenon of interest. Even univariate repeated measures data can be analyzed by employing univariate or multivariate data analytic strategies. Both univariate and multivariate methods can be valuable under certain conditions and when various assumptions are met. The present paper examined both univariate and multivariate approaches to analyzing repeated measures data and compared the results of these methods with classical ANOVA and multiple regression analyses. A small heuristic data set was utilized to make the discussion concrete and to facilitate conceptual understanding of the material.

The primary objective of all scientific research is to gather new information about a phenomenon of interest and to convey to other interested parties conclusions in an effort to further the accumulation of scientific knowledge. Since science is concerned primarily with repeatable and replicable experiments, researchers are required to examine and reduce the influence of random effects that might contaminate results and consequently distort conclusions. One manner in which researchers have sought to reduce the influence of random effects is to utilize a separate set of individuals for different levels of a treatment condition (termed a between-subjects design). The logic employed is that any individual differences manifested prior to the implementation of the treatment can be alleviated by randomly assigning individuals to only one of the various conditions. This approach has the advantage of being less cumbersome to participants but frequently places a strain on the researcher, as it is necessary to amass larger numbers of participants for each treatment and control condition so that the random assignment mechanism can work effectively to minimize the influence of outliers (Girden, 1992).

Another way to control for the influence of individual differences is through the use of the same group of individuals measured repeatedly over time, occasions, or conditions (often termed repeated measures or within-subjects designs). In this manner each individual serves as his or her own control group, and the aberrant influences garnered by one individual on a single measurement will likely be manifested by that individual on subsequent measurements as well. Consequently, the influence of different levels of the treatment condition, as well as the influence of intrapersonal (within-subjects) and interpersonal (between subjects) differences, can be examined. Advantages afforded by this approach include a decreased number of participants required to achieve reasonable power against Type II error and distorted effects due to outliers (Stevens, 1996).

The differences between the two methodologies previously described can be illustrated in a brief example. Suppose a pharmaceutical company has developed a new drug for adult depression and desires to test the drug on a group of human participants for any potentially harmful side effects. The company, having just developed the drug treatment and thus nearly exhausting the financial resources reserved for the project, needs to find a cost effective manner in which to test the treatment. One viable alternative, the between-subjects approach, would be to gather a large group of individuals and randomly assign participants to six different dosage conditions ranging from a placebo treatment (sugar pill) to a 450mg treatment of the medication (spaced in 150mg increments). Supposing a minimum number of 20 participants are required per condition, the drug company would need to recruit a minimum of 120 individuals to participate in the study.

Another potential option would invoke a repeated measures design in exploring the differential effects of the drug treatment. In this approach a group of individuals would also be gathered, but each individual would receive each of the six levels of the drug treatment in a random order. Again supposing a minimum of 20 participants per condition, a total of only 20 participants would be required to complete the study. Consequently, the demands on participants is greater in the repeated measures design, but the benefits of employing such a design can be both statistically elegant and satisfying for the researcher.

Because employing repeated measures designs can be both practical and powerful in social science research (Girden, 1992), it is lamentable that more researchers have not opted to utilize this type of design. One reason for the under utilization of this approach is disagreement regarding the appropriate analytic technique to use to evaluate results, as either a univariate or multivariate approach can be invoked (Algina & Keselman, 1997; Girden, 1992; Keselman, Keselman, & Lix, 1995; Keselman, Lix, & Keselman, 1996; Maxwell & Delaney, 1990). As noted by several authors (Algina & Keselman; Edwards, 1985; Girden; Maxwell & Delaney), each analysis has distinct advantages and disadvantages, and each type of analysis will provide a more powerful result under certain conditions and when certain statistical assumptions are satisfied.

The purpose of the present paper is to examine the similarities and differences in the univariate and multivariate analysis of univariate repeated measures data involving repeated measurements of a single dependent variable. The discussion will include comparisons with classical ANOVA and multiple regression approaches to repeated measures analysis. Power and practicality of the various approaches will be illustrated by utilizing a small heuristic data set throughout the paper.

Univariate Approach to Repeated Measures Analysis

The univariate approach to repeated measures analysis, often termed a repeated measures ANOVA, invokes many concepts also employed in a classical ANOVA analysis in that the sums of squares (*SS*) is partitioned into various constituent components. In repeated measures ANOVA, however, it is possible to further partition variation that classical ANOVA simply terms 'error' variance. Thus, in an effort to understand the advantages afforded by invoking repeated measures designs, it is first necessary to review some pertinent components of classical ANOVA.

Mechanics and Logic Underlying Classical ANOVA

ANOVA (Fisher, 1925) is a statistical procedure utilized to compare means ($k \geq 2$) in an effort to determine if the means differ from one another (Edwards, 1979). In the most traditional application of ANOVA, researchers typically utilize the technique to test the statistical significance of the differences among or between groups of means, but ANOVA can also be used to generate variance-accounted-for effect sizes as well (Wilcox, 1987). As stated by Shavelson (1988), "The purpose of... ANOVA is to compare the means of two or more groups in order to decide whether the observed differences between them represent a chance occurrence or a systematic effect" (p. 342).

Classical ANOVA with One Between Factor

A construct crucial to understanding ANOVA is the most basic unit of all statistical analysis: variance. As stated by Haase and Thompson (1992), "variance is the 'stuff' on which all analysis is based" (p. 3). Variance is the degree to which the scores are spread out or dissimilar in a data set. As the name implies, analysis of variance is concerned primarily with determining the sources of variation present within a set of data.

The most simplistic manner in which to examine the shared variance or the differences manifested in group means is by examining distinct groups on a singular independent variable, commonly referred to as a one-way ANOVA. In this context the word "way" (also termed factor) refers to the singular independent variable under investigation (it is assumed in ANOVA that only one dependent variable will be utilized). Researchers employing the one-way ANOVA technique typically have a group of individuals, which has been either naturally or systematically divided into two or more subgroups on a singular classification scheme. The objective of using the one-way ANOVA is to determine if the groups under investigation are appreciably different from one another by ascertaining what proportion of the total dependent variable variance is accounted for by each group.

The process of executing a one-way ANOVA involves consulting several statistics utilized in the computational process. Among the first statistics derived in this process is (a) the sum of squares total (SS_T) - the total variation in the scores on the dependent variable; (b) the sum of squares between groups (SS_B) - the variation in the scores on the dependent variable attributed to the independent variable; and (c) the sum of squares within groups (SS_w) - the variation in the scores on the dependent variable that cannot be accounted for by the independent variable. An interesting property of the three *SS* components is that in a balanced design SS_B and SS_W *always* sum to SS_T, as SS_B and SS_W are partitioned areas of SS_T. A useful metaphor in conceptually understanding *SS* is to equate SS_T with an entire pie. The pie is constituted of two individual slices, one representing SS_B and one representing SS_W. Thus, SS_B and SS_W are always perfectly uncorrelated and are completely separate and unique entities. The size of each slice of pie corresponds to the proportion of variance accounted for by each *SS* component.

ANOVA Test Statistic and Decision Process

The test statistic used in ANOVA to determine if differences are present in the analyzed data is called the F_{ratio} or $F_{\text{calculated}}$ (F_{calc}). This statistic is generated by dividing the mean square between (SS_B divided by the degrees of freedom between (df_B)) by the mean square within (SS_W divided by degrees of freedom within (df_W)). The resultant F_{calc} can be compared to the critical values of the $\mathcal{F}$-distribution (F_{crit}) contained in most statistical texts, and if the F_{calc} supersedes the F_{crit} value (located by using the degrees of freedom between and the degrees of freedom within) at the specified α level, then the result of the analysis is said to be "statistically significant." A result that is statistically significant implies that given the sample size and the specified α level, the researcher has decided to reject the null hypothesis. A statistically significant result, however, does not imply that the result is inherently important (Thompson, 1993; 1996a; 1999a; 1999b). Conversely, an F_{calc} that fails to exceed the F_{crit} value is subsequently considered not statistically significant but is not necessarily regarded as inherently unimportant, as statistical significance does not furnish any information to a researcher in regard to result importance (Thompson, 1999a, 1999b).

Another manner in which to render a decision in ANOVA is through the computation of a statistic called eta^2, a variance accounted for effect size functionally equivalent to R^2. Eta^2 is an uncorrected effect size (see Snyder & Lawson, 1993 for a full treatment of effect sizes) and is computed by dividing SS_B by SS_T. The resultant statistic informs the researcher as to what portion of the variance can be explained with knowledge of to which group the participants belonged.

Multiple Between-Subjects Factors in Classical ANOVA

In a two-way ANOVA, two independent variables are examined in relation to a singular dependent variable. By employing this analysis, it is possible to examine the effects of one or both of the ways on the dependent variable as well as the effects of combinations of the ways. By affording researchers the ability to examine the effects of a combination of the ways on the dependent variable, ANOVA becomes a very powerful tool in discerning the relationships between different variables.

As stated previously, ANOVA partitions the variance on the dependent variable into two distinct components, SS_B and SS_W. Recall from the earlier discussion of the one-way case that each partition is a unique and separate portion of the SS_T. In the two-way case, the SS_B portion is further partitioned into two components, the *SS* associated with the main effects and the *SS* associated with the interaction. For a balanced design, each portion of SS_W is still considered a separate and unique portion of the SS_T and will *never* overlap with any other *SS* partition. A final *SS* partition splits the *SS* allotted to the main effects into one portion associated with the "A" way main effect and one portion associated with the "B" way main effect. Thus, in the factorial two-way case, there are four *SS* partitions resulting in the components of $SS_{\text{A-way}}$, $SS_{\text{B-way}}$, $SS_{\text{INTERACTION}}$, and SS_W. Consequently, in a balanced design, any information generated by one *SS* component does not render any information about another *SS* component, as all the *SS* partitions are *perfectly uncorrelated* (i.e., orthogonal). The only exception to this dynamic is when one partition equals SS_T, because the other partitions must then equal zero.

The resultant effect of performing a multi-way classical ANOVA analysis can be either positive or negative (Benton, 1991). Two dynamics are at work when discerning the effect of multiple ways on a given analysis: the extent to which the SS_W and df_W are reduced. Reducing the SS_W without altering the associated df_W reduces the MS_W component and provides a larger F_{calc} value. Unfortunately, partitioning more variance in classical ANOVA costs the researcher, and the remuneration for partitioning more variance is a reduction in the df_W. The effect of reducing the df_W can counteract a reduction of the SS_W, as reducing the df_W will increase the MS_W and will subsequently reduce the F_{calc} value. Thus, as in any analysis, there is a cost-benefits decision to be rendered when including more than one factor in a classical ANOVA. If the newly included factor consumes a substantial portion of SS_W without appreciably reducing the df_W, the resultant F_{calc} value is more likely to be statistically significant.

Single Within-Subjects Factor Designs

Invoking a single within-subjects factor (e.g., repeated measures) design allows the researcher to partition variance on the dependent variable in much the same way as in classical ANOVA. The difference is that in a single factor repeated measures design, the total variation is partitioned into that component associated with variation due to individual differences in the participants (SS_{SUB}) and variation due to differences in the levels of the treatment condition (SS_{TRT}). The remaining variation (typically termed SS_{RES}) is often considered an interaction component because it represents variation due to the unique combination of the participants and the treatment levels. As noted previously, in classical ANOVA with a single between-subjects factor, it was only possible to partition variance into between-groups (variation due to differences in groups of participants) and within-groups variation (error). Because repeated measures analysis explains more of what classical ANOVA simply terms error, it is possible to reduce the MS_{RES} value and subsequently generate a larger F_{calc} value with fewer participants (Girden, 1992). Consequently, repeated measures ANOVA designs tend to be more efficient because fewer participants are required to conduct an analysis that may generate more statistical power than classical ANOVA (i.e., between-subjects) designs.

Classical ANOVA includes individual differences in performance as error (SS_{W}) because with only one measurement of each person the variance attributable to each person as a unique individual cannot be estimated. In repeated measures analyses, however, the variance due to differences in people can be estimated. Thus, in a single within-subjects factor situation, variance is partitioned into between-subjects (differences across the different treatment effects), within-subjects (differences due to intrapersonal variation), and residual variation.

Carry-Over, Latency Effects and Counterbalancing

As stated previously, repeated measures designs require the same group of participants to be measured on multiple occasions. It has been recognized in the social sciences that the order of presentation of stimuli can sometimes have a differential effect on participants' responses as some experiments involve repeated exposure to the same task (Girden, 1992; Keppel, 1991; Keppel & Saufley, 1980; Keppel & Zedeck, 1989; Stevens, 1996). As such, carry-over and latency effects are common problems in repeated measures research. A latency effect refers to a situation in which the effect of a treatment is not evident until a subsequent level of the treatment is introduced. A latency effect may predispose a researcher to erroneously contend that the administered treatment had little to no effect on the monitored behavior when, in actuality, the effect of the treatment was not evidenced until an additional condition had been implemented. Similarly, a carry-over effect refers to the influence of a previous level of treatment on the observed behavior in a subsequent level of the same treatment condition.

Carry-over and latency effects tend to skew results by influencing the responses of participants and can be both positive or negative in nature. A strategy termed *counterbalancing* is frequently employed in repeated measures research to help combat carry-over and latency effects. Counterbalancing involves presenting levels of a treatment condition so that each level occurs equally often at each stage of practice and so that each level precedes another level as many times as it follows the level. When treatment administrations are counterbalanced, it is possible for a researcher to discern if certain combinations of treatment levels adversely affected the observed results.

Counterbalancing the presentation of treatment stimuli in repeated measures designs is critical to the generation of data that accurately represents the effect of a given treatment on the behavior of interest. The following paradigm can be invoked when determining the presentation order of the stimuli provided there are an even number of treatment levels and the number of participants is some multiple of the number of conditions:

$$1, 2, n, 3, n\text{-}1, 4, n\text{-}2, 5, n\text{-}3, 6, n\text{-}4, \text{etc.}$$

Table 1. Counterbalancing Order for Design with Six Treatment Conditions

	Trial Number					
Person	One	Two	Three	Four	Five	Six
A	1	2	6	3	5	4
B	2	3	1	4	6	5
C	3	4	2	5	1	6
D	4	5	3	6	2	1
E	5	6	4	1	3	2
F	6	1	5	2	4	3

Note. Each number (1-6) indicates which level of the treatment the participant would receive on each corresponding trial.

If six treatment levels were to be administered in a given study, the first participant would be presented the treatments levels in the order of 1, 2, 6, 3, 5, 4. To derive the order of presentation for the second participant, it would be necessary to add a 1 to each of the numbers in the preceding order:

$$2=(1+1),\ 3=(1+2),\ 1\equiv\ (1+6 \text{ reduces to } 1),\ 4=(1+3),\ 6=(1+5)\ 5=(1+4).$$

Using this generation rule, the completed order of presentation for five participants when administered five treatment conditions is presented in Table 1.

Notice that each treatment condition is presented before and after every other treatment condition. In the case that there is an odd number of levels of the treatment condition, the first order of presentation is derived as previously illustrated but the second order is computed by reversing the sequence of the first presentation. For example, if there were three levels of treatment to be administered, the first order of presentation would be 1, 2, 3 and the second order of presentation 3, 2, 1. The third order of stimuli presentation would be 1≡ (1+3, reduces to 1), 3=(1+2) and 2=(1+1). This procedure would be repeated until all of the participants were assigned stimuli presentation orders. The examples presented here conform to the criteria delineated by Girden (1992), as each treatment level occurs equally often at each stage of practice and precedes as many times as it follows a level.

The concept of counterbalancing, although important and necessary in repeated measures analyses, cannot always remedy some of the problems experienced in employing repeated measures designs. In some instances the order of presentation of the stimuli is essentially irrelevant in that after the stimuli are presented on the first occasion the participant might be able to demonstrate a substantial practice effect on later trials, which skews the responses on the dependent variable. Consider, for example, a researcher teaching undergraduate students the names of five important psychologists, and then examining the effects of differential drug treatments on memory recall. Many, if not all, of the participants may successfully commit the names to memory on the first trial and will then be able to recite them after each of the treatments regardless of the order of presentation. Counterbalancing is important in generating accurate data in repeated measures designs, but often even a good correction method cannot repair a poor research design.

SS Partitions and the Test Statistic

The remainder of the present paper uses a singular heuristic example to examine the various statistical and conceptual properties of univariate and multivariate approaches to repeated measures analyses. Consider the following situation: A pharmaceutical company has developed a new prototypical drug that is purported to alleviate depressive symptomalogy in human adults. In a bid to gain FDA approval of the drug so that the treatment can be disseminated in the public sector, the company tested the medication on a small sample of human participants. The drug was presented in each of four dosage levels (0mg, 150mg, 300mg and 450mg) to each of five individuals. The different levels of treatment were counterbalanced across participants, and each additional level was given only after a sufficient time had passed to allow traces of the previous treatment to exit the participants' systems. After each dosage had an opportunity to take effect, the participants were administered a depression inventory (i.e., dependent variable) to assess

Table 2. Hypothetical Data Matrix for Medication Dosage Study

	Trial Number					
Subject	0mg	150mg	300mg	450mg	Sum	Mean(Y_i)
1	1	10	14	18	43	10.75
2	3	6	15	19	43	10.75
3	3	8	11	20	42	10.50
4	4	8	13	16	41	10.25
5	5	9	13	18	45	11.25
ΣY_j	16	41	66	91	214	
Mean(Y_j)	3.2	8.2	13.2	18.2		10.70

Note. Scores on the hypothetical depression inventory range from 0 ‘critically depressed’ to 25 ‘not depressed’.

their level of depression (ranging from 0 ‘critically depressed’ to 25 ‘not depressed’). After the conclusion of the last treatment, a data matrix was compiled and examined using a single within-subject factor (drug treatment) repeated measures analysis. These hypothetical data are presented in Table 2.

The first step in completing the repeated measures ANOVA is to partition the variance on the dependent variable. As mentioned earlier, the variance will be decomposed into three main components: A portion associated with the differences in the participants (SS_{SUB}); a component associated with differences in the treatment conditions or intervals (SS_{TRT}); and a portion associated with error variance and the effects of individual participant differences across treatment conditions (SS_{RES}). Thus, the decomposition of the total variation on the dependent variable can be described in the equation:

$$SS_{\text{T}} = SS_{\text{SUB}} + SS_{\text{TRT}} + SS_{\text{RES}}\ .$$

The SS_{T} component represents the sum of the squared deviation of each dependent variable score from the grand mean. SS_{SUB} is computed by summing the squared deviations of each subject mean from the grand mean. SS_{TRT} represents the sum of the squared deviations of the treatment means from the grand mean. Finally, SS_{RES} is computed by subtracting SS_{SUB} and SS_{TRT} from SS_{T}. Thus, for the present example, the equation can be written:

$$660.20 = 2.20 + 625.00 + 33.00.$$

The next step is to compute the appropriate degrees of freedom for each source of variation. The formulas for the *df* calculations are presented in Table 3. As noted earlier, the residual term is considered an interaction effect (subjects by treatment intervals) and thus the *df* for this component is calculated by multiplying the *df* for subjects (df_{SUB}) by the *df* for treatment intervals (df_{TRT}). After calculating the appropriate *df*, the mean squares, F_{calc}, eta^2, and $omega^2$ values are calculated. Notice that the only F_{calc} value that is computed is the value corresponding to the treatment interval source of variation because it is usually of primary interest.

The results of the repeated measures ANOVA on the example data are very favorable. The different treatment conditions accounted for 94.67% of the total variation on the depression scores and rendered a statistically significant result at the $\alpha = .05$ level. Additionally, there was very little variation in the performance of individual participants when averaged across conditions as illustrated by the very small *SS* value for the between subjects source of variation (2.20). Based on the effect size and statistical significance of these results, the pharmaceutical company can report that different dosages of the anti-depression medication produced a decrease in the overall depression scores of the participants involved in the study.

A comparison of the results in the repeated measures ANOVA with the results generated by a classical ANOVA of the same set of data produces an interesting topic for discussion. In order to invoke a classical ANOVA with this data, it would be necessary to measure each individual only once as classical ANOVA examines effects between subjects; consequently, 20 participants would be required to complete the exact same study. In the classical ANOVA approach rather than each participant receiving all four

Table 3. Summary Table for Repeated Measures ANOVA with Table 2 Data.

Source	*SS*	*df*	*MS*	F_{calc}	*eta*^{2}	*omega*^{2}
Subjects	2.20	*n*-1 = 4	0.55			
Intervals (TRT)	625.00	*k*-1 = 3	208.33	75.76	0.9467	0.9144
Residual	33.00	(*n*-1)(*k*-1) = 12	2.75			
Total	660.20	(*n*)(*k*)-1 =19				

Note. *n* = number of Subjects, *k* = number of Treatment Intervals.

omega^2 = $(SS_{\text{TRT}} - ((k\text{-}1)MS_{\text{RES}}))/(SS_{\text{T}} + MS_{\text{SUB}} + (nMS_{\text{RES}})$.

levels of treatment, each individual would be randomly assigned one level of treatment. Consequently, it is only possible to examine the differences between the levels of treatment and not the differences between individual participants. Thus, the variance on the dependent variable is partitioned into one less component in classical ANOVA than in the repeated measures analysis.

The results of the four level one-way ANOVA on the example data are presented in Table 4. Notice that the SS_{RES} (error) is larger in the classical ANOVA analysis (i.e., SS_{W}) because there is one less variance partition. Also note that the df_{RES} is smaller in the repeated measures ANOVA than in the classical ANOVA. Because only 2.20 *SS* units were accounted for by the inclusion of this source of variation in the repeated measures analysis and the cost of examining that piece of information was 3 *df* from a small number of 20 total observations, employing a repeated measures analysis in this case could have been detrimental to the outcome of the study in terms of statistical significance due. Even though both results are statistically significant at the $\alpha = .05$ level, the F_{calc} value is much larger in the classical ANOVA analysis. All things being equal, larger F_{calc} values will lead to a better chance of obtaining statistical significance, if the researcher is concerned with doing so.

Notice, however, that the variance accounted for by the treatment component (94.67%) remains identical in each analysis. In a real world setting with actual data and a larger sample size, the distinction between these two approaches would be more pronounced, and the power of utilizing a repeated measures approach could be illustrated more convincingly. Remember, however, that very similar results were generated in both analyses, but the repeated measures approach only utilized five participants total compared to the 20 individuals required to complete the classical ANOVA analysis.

Assumptions in Univariate Repeated Measures Analysis

Invoking a repeated measures ANOVA, however, does not come without several difficulties and considerations. As stated by Girden (1992, p. 13),

> . . . several assumptions . . . were recognized by R.A. Fisher in the 1940's but were not demonstrated until the 1950's . . . with the result that many earlier studies involving repeated measures may have reached erroneous conclusions regarding the effect(s) of the independent variable(s).

It is difficult to understand why researchers and statisticians did not better explore and develop the statistical assumptions of repeated measures designs (i.e., compound symmetry and sphericity) earlier in the history of their use. Repeated measures were being used shortly after the development of ANOVA in 1925, and countermeasures were developed to contend with treatment carry-over, latency and practice effects long before Box (1954) articulated the problem of departures from compound symmetry.

Compound Symmetry

The term, *repeated measurements*, implies that each individual is measured on more than one occasion. Consequently, Box (1954) contended that the assumption of compound symmetry, which states that it is necessary for the variances and covariances of different treatment levels to be equal, must be satisfied for the results of the repeated measures analysis to be valid. The observations generated by each individual are independent of the observations produced by any other participant, but each individual's score on a treatment level is often linearly dependent on or correlated with the previous scores. The degree of correlation between a given individual's scores over the treatment conditions affects the

Table 4. Results of Four-Level One-Way ANOVA with Table 2 Data.

Source	*SS*	*df*	*MS*	F_{calc}	*eta*^{2}
Between	625.00	*k*-1 = 3	208.33	94.70	.9467
Residual (Within)	35.20	*k(n*-1) = 16	2.20		
Total	660.20	(*n*)(*k*)-1 =19			

statistical significance of the repeated measures result because the MS_{RES} will decrease as the degree of correlation between the observations increases (thus leading to a smaller MS_{TRT} and a smaller F_{calc} value).

It is important, therefore, to examine the degree of covariation among each pair of treatment scores (Algina & Keselman, 1997; Girden, 1992; Maxwell & Delaney, 1990), as these scores must be equal when invoking an analysis that computes an F_{calc} value (MS_{TRT} / MS_{RES}) with (*k*-1) and (*k*-1)(*n*-1) *df.* Prior to Box (1954), researchers and statisticians focused solely on the equality of the variances (homogeneity of variance assumption) and ignored any influence that unequal covariances might render on the results of a repeated measures analysis. If the covariance between two sets of scores is defined as

$$\text{COV}_{xy} = \sum (X - \bar{X})(Y - \bar{Y})/(n\text{-}1) \text{ ,}$$

then the formula for the covariation of a set of scores with itself could be written as

$$\sum (X - \bar{X})(X - \bar{X})/(n\text{-}1) \text{ .}$$

The latter of the two formulas is the formula for the variance of a set of scores, thus indicating that variance is nothing more than the degree of covariation of a set of scores with itself. Thus, Box (1954) noted that examining the equality of variances in repeated measures analyses was only half of the larger issue; because variance can be defined as the covariation of a set of scores with itself, it is necessary to also examine the equality of covariances of all pairs of the treatment levels as well.

The assumption of compound symmetry derives its name from a matrix that contains information about both variances and covariances of a set of scores. For a given set of scores, the variance of each set of scores and the covariances between all possible pairs of the scores can be arranged in a single matrix termed a variance-covariance matrix. This is a special matrix of rank *k*, where *k* indicates the number of treatment levels in the single within factor repeated measures analysis. The variances of the *k* treatment levels are presented on the main diagonal of the matrix and the covariances between each pair of treatment levels are presented on the off diagonals. A variance covariance matrix that exhibits equal variances and equal covariances is said to demonstrate compound symmetry.

Violating Compound Symmetry

In addition to indicating that the equality of covariances between pairs of the treatment conditions must be considered, Box (1954) noted other important considerations as well. Box effectively demonstrated that if an F_{calc} value does not originate from an $\mathcal{F}$-distribution with (*k*-1) and (*k*-1)(*n*-1) *df*, then it is not a part of that $\mathcal{F}$-distribution. Box contended, therefore, that such an F_{calc} value belongs to an $\mathcal{F}$-distribution that is corrected by a factor called epsilon (ε) and which results in ε(k-1) and ε(k-1)(n-1) degrees of freedom. He further noted that the correction factor invoked by the epsilon value is more severe as the covariances become more unequal and is approximately equal to 1.0 when the equality of covariances is demonstrated.

The implication on the statistical significance of a given F_{calc} value can be profound when the epsilon correction factor is severe. Consider the following example: A researcher performs the univariate ANOVA for a single within-subjects factor and generates an F_{calc} value of 3.15 after performing all relevant calculations. The researcher then consults a table of critical values from the $\mathcal{F}$-distribution to determine if the result is statistically significant. Upon examination, the researcher learns that the F_{crit} value for the study, *F*(4, 19) at the $\alpha = .05$ level, is 2.90. The researcher then revels in the limelight of a statistically significant result generated by varying the treatment conditions.

This result, however, is only valid provided that the variances and covariances of the treatment levels are equal. If the variances and covariances between treatment levels in the example study are not equal, then the researcher must alter the *df* with which to evaluate the statistical significance of the F_{calc} value by a factor of epsilon. If the epsilon value is equal to .5, the new *df* for the F_{calc} would be .5(4) and .5(19); thus, when accounting for the correction factor, the *df* of 2 and 10 must be used. When the F_{crit} value using these *df* is examined, the researcher is dismayed to learn that the corrected F_{crit} value is 4.10, thus rendering the F_{calc} value of 3.15 no longer statistically significant at $\alpha = .05$. The conceptual point here is that when epsilon provides a severe correction for the inequality of variances and covariances, the *df* for the F_{crit} value can be radically affected. Consequently, researchers may use inappropriate *df* that generate a F_{crit} value that is smaller than it should be and by which a researcher may reject a null hypothesis when it should not be rejected.

Box (1954) demonstrated that the upper bound for the epsilon correction is 1.0 which indicates that the epsilon corrected *df* are exactly equal to the $\mathcal{F}$-distribution *df*. At this extreme a correction factor is not invoked and erroneous conclusions are avoided. Geisser and Greenhouse (1958) determined the lower bound for the epsilon correction factor by positing that the lower bound of epsilon was equal to $1/(k\text{-}1)$. In the previous example with five treatment levels administered to five subjects, the lower bound of epsilon would have been 1/(4-1) or .33. At this extreme the most severe correction would be invoked, and the *df* for F_{crit} would be exactly one third of their original magnitude.

Sphericity

Research conducted after Box's (1954) work on compound symmetry indicated that although compound symmetry is important to consider in repeated measures analyses, it is not a necessary condition for conducting them (Girden, 1992). Sphericity (sometimes called circularity) is considered a necessary and sufficient condition to conduct repeated measures analyses (Huynh & Feldt, 1970; Rouanet & Lepine, 1970). Sphericity is the degree to which variances in the differences between pairs of treatment scores are equal. The notion of sphericity is a more flexible assumption and subsumes compound symmetry as a special case. That is, the variances of differences between treatment levels would be equal when the variances and covariances were all equal.

To satisfy the assumption of sphericity, the variances of the differences in all pairs of treatment combinations must be homogeneous (e.g., the variance of level 1 and 2 must equal the variance of level 2 and 3, etc). The variance of a difference between two treatment conditions (1 and 2) can be defined as

$$\sigma^2(Y_1 - Y_2) = \sigma^2_1 + \sigma^2_2 - 2\sigma_{12}$$

where σ^2_1 is the variance of on a set of scores, σ^2_2 is the variance of another set of scores, and σ_{12} is the covariance of the two sets of scores.

Assessing the equality of variances of differences in the treatment levels can be illustrated using the heuristic example data presented in Table 2. Table 5 contains the variances and covariances for each of the treatment levels. By invoking the formula described above for the variance of a difference between treatment levels, it is possible to compute the variance of the difference between treatment levels 1 (0mg) and 2 (150mg). The variance of the difference would be

$$\sigma^2(Y_1 - Y_2) = 2.1998 + 2.1998 + 2(-.5500) = 3.2996.$$

Because the variances and covariances of each of the treatment levels are exactly equal in this example, all of the variances of difference would be equal to 3.2996. The assumption of sphericity, therefore, would be satisfied in this example. This example has the additional feature of being compound symmetrical, as all of the variances and covariances are exactly equal.

Another way to assess the sphericity of a data set is to examine the matrix of orthonormal contrasts (Stevens, 1996). If the multivariate identity

$$\mathbf{C}'\textstyle\sum\mathbf{C} = \sigma^2\mathbf{I}$$

where $\mathbf{C}$ is a matrix of $(k\text{-}1)$ orthonormal contrasts, $\mathbf{C}'$ is the transpose of $\mathbf{C}$, $\sum$ is the variance-covariance matrix and $\sigma^2\mathbf{I}$ is an identity matrix (with equal variances on the main diagonal and zeros on the off diagonal) is true, the assumption of sphericity is satisfied. The first step in assessing sphericity in this manner is to create a set of orthogonal contrast variables. One set of orthogonal contrast variables for the example data are presented in Table 6. Notice that there are three contrast variables present corresponding to the $(k\text{-}1)$ contrasts possible. The contrast variables are then normalized by invoking a multiplicative constant such that the sum of the squared transformed coefficients in a given contrast is equal to 1.0. This

Table 5. Correlation and Variance-Covariance Matrix for Table 2 Data

Tx	0mg	150mg	300mg	450mg
1	2.20	-0.55	-0.55	-0.55
2	-0.55	2.20	-0.55	-0.55
3	-0.25	-0.25	2.20	-0.55
4	-0.25	-0.25	-0.25	2.20

Note. Variances are on the main diagonal. Correlations are below and Covariances are above the main diagonal.

Table 6. Orthogonal Contrast Variables.

Tx	C_1	C_2	C_3
1	1	1	1
2	-1	1	1
3	0	-2	1
4	0	0	-3

Table 7. Orthonormal Contrast Matrix **C**.

Tx	C_1	C_2	C_3
1	.707	.408	.289
2	-.707	.408	.289
3	0	-.816	.289
4	0	0	-.866

Note. Tx = treatment level; C_1 = contrast variable 1; C_2 = contrast variable 2; C_3 = contrast variable 3

is accomplished by first squaring the coefficients in the contrast and dividing by the derived number. For example, the second contrast variable contains the contrast coefficients of 1, 1, and -2. It is necessary to first square each coefficient, $(1)^2 + (1)^2 + (-2)^2 = 6.00$, and then to divide each coefficient by the result, 1/√6, 1/√6, and (-2/√6). This procedure would be repeated until all contrasts are transformed. The orthonormalized variables are placed in a matrix as presented in Table 7. The product of the equation, $\mathbf{C}'\sum\mathbf{C} = \sigma^2\mathbf{I}$, is then computed to determine if the assumption of sphericity is met (Stevens, 1996).

Violating the Assumption of Sphericity.

Violating the assumption of sphericity can have the same grievous consequences as violating the assumption of compound symmetry. If the variances of the differences in levels of the treatment conditions are not equal, the *df* for the F_{calc} value will be smaller than normal by a value of ε. Thus, just as violating compound symmetry caused the F_{crit} value to increase in magnitude, so too, does violating the assumption of sphericity. If sphericity is not examined, the unwary researcher will tend to reject the null hypothesis more often than it should be rejected resulting in higher Type I error rates (Stevens, 1996).

Correcting for Violations in the Sphericity Assumption

Adjusted Degrees of Freedom. Correcting for a violation in the sphericity assumption involves the computation of the epsilon parameter previously described. As noted earlier, Box (1954) defined the upperbound of epsilon to 1.0, and Geisser and Greenhouse (1958) calculated the lower bound of epsilon to be 1/(*k*-1), where *k* is the number of treatments utilized in the study. To find the actual value of epsilon, the following formula generated by Greenhouse and Geisser (1959) must be invoked,

$$\varepsilon = \frac{k^2(\bar{s}_{ii}^2 - \bar{s}_{**})^2}{(k-1)(\sum\sum s_{ij}^2 - 2k\sum \bar{s}_i^2 + k^2\bar{s}_{**}^2)}$$

where $\bar{s}_{ii}^2$ is the mean of entries on the main diagonal of the variance-covariance matrix, $\bar{s}_{**}$ is the mean of all entries in the variance-covariance matrix, s_{ij} is the ij^{th} entry of the variance-covariance matrix and $\bar{s}_i$ is the mean of all entries in the row *i*. The calculation of epsilon can be illustrated using the example data in Table 2. The calculations

$$\varepsilon = \frac{4^2(2.2000 - 0.1375)^2}{(3)(22.9900 - (8)(0.075625) + (16)(0.1375^2))}$$

result in an ε value of 1.0. This value indicates that the sphericity assumption is perfectly met in the example data and that no corrections to the degrees of freedom are required. To determine if the result is statistically significant, the researcher would simply use the same degrees of freedom as calculated when constructing the summary table.

Table 8. Hypothetical Data Matrix for Worst Case of Violating Sphericity Assumption.

	Trial Number					
Subject	0mg	150mg	300mg	450mg	Sum	Mean(Y_i)
1	1	10	15	20	46	10.50
2	2	6	11	16	35	8.75
3	3	7	12	17	39	9.75
4	4	8	13	18	43	10.75
5	5	9	14	19	47	11.75
ΣY_j	15	40	65	90	210	
Mean(Y_j)	3.0	8.0	13.0	18.0		10.70

Consider an example in which the epsilon value invokes the greatest possible correction factor on the degrees of freedom used to determine F_{crit}. If the example data in Table 2 are changed only slightly to resemble the data in Table 8, the results of the repeated measures analysis are radically different. The correction factor invoked in the Table 8 data is $\varepsilon = .33$ or the minimum value possible in a design with four treatment levels (1/*k*-1) because the scores are linearly dependent and completely violate the sphericity assumption. Rather than using (k-1)=3 and (k-1)(n-1)=12 degrees of freedom to arrive at an F_{crit} value of 3.49 (α = .05), each value would be corrected by the value of ε. Thus, the new degrees of freedom for the F_{crit} value would be 1 and 4 and would render an F_{crit} value of 7.71 (α = .05). The results in the present study would still be statistically significant, but in situations where the F_{calc} value is only slightly larger than the uncorrected F_{crit} value, even a small correction by epsilon can alter the results from statistically significant to not statistically significant. This comparison allows the power of a repeated measures design to be illustrated: If ε is equal to 1.0, the repeated measures ANOVA demonstrates the power equivalent to (*n*x*k*) participants. Consequently, fewer participants can be utilized without compromising sufficient power to reject the null hypothesis.

Estimates of Epsilon. There are several estimates of ε currently available, and each has its own distinct advantages and disadvantages (Maxwell & Delaney, 1990; Stevens, 1996). The most commonly employed indices of ε are the Greenhouse-Geisser ε (Greenhouse & Geisser, 1959) and the Huynh-Feldt ε (Huynh & Feldt, 1970). Because both ε values are computed in most statistical software packages, it is important to understand the particular bias that each estimate produces.

When there is only one sample, as with the present example, the Greenhouse-Geisser and Huynh-Feldt estimates are identical. However, for more two or more independent groups (i.e., split-plot design), the Greenhouse-Geisser ε tends to underestimate the true value of epsilon across the range of values, but the underestimation is even more pronounced as epsilon approaches 1.0. Consequently, the Greenhouse-Geisser ε will produce a very conservative estimate of the *df* utilized to obtain the F_{crit} value which may result in not rejecting the null hypothesis as often as might be indicated by the data. Conversely, the Huynh-Feldt ε produces an overestimation of the true value of ε and may result in a smaller F_{crit} value than is indicated by the data. Thus, when using the Huynh-Feldt ε, researchers may reject the null hypothesis more often than they should. Because the two most popular estimates of ε produce biased results, authors have recommended averaging the two indices as a more accurate estimate of ε (Barcikowski & Robey, 1984; Girden, 1992; Stevens, 1996). If one must be chosen over the other, however, it is always somewhat safer to utilize the Greenhouse-Geisser ε as it produces a more conservative correction factor.

Guidelines. Guidelines have been presented to facilitate the correct interpretation of univariate repeated measures analyses (Greenhouse & Geisser, 1959). Due to the advent of statistical software that readily and painlessly computes many different estimates of ε, these guidelines have limited pragmatic value but warrant brief consideration. As provided by Girden (1992, p. 21) the steps are as follows:

1. Compare the obtained [F_{calc}] with the tabled value corresponding to [k-1] and [k-1][n-1] *df*. If it is not greater than this most liberal value, stop at this point. It will not be significant when degrees of freedom are reduced.

2. If the obtained [F_{calc}] is significantly higher than the most liberal value, enter the table with 1 and [n-1] *df*. If the obtained F is greater than this most conservative value, it is significant. Stop at this point.
3. If the obtained [F_{calc}] is higher than the tabled value for df = [k-1] and [k-1][n-1], but lower than the tables value for df = 1 and [n-1], then the ε adjustment should be applied.

It is not possible in the present example to illustrate this dynamic because the ε value of the data set is 1.0, but interested readers can examine the analysis of the data set in Girden (1992) for a good example of a situation where these guidelines might prove helpful.

Another Univariate Approach to Repeated Measures Analysis

As mentioned previously, it is possible to employ multiple regression analysis to examine the effects of repeated measures data. A brief treatment of this analytic technique is presented here, and interested readers are referred to Edwards (1985) for a more detailed discussion.

To utilize this approach it is necessary to first construct k contrasts (where k represents the number of treatment levels) in which each contrast variable represents a separate effect. In the example data set four treatment conditions were utilized; thus, four orthogonal contrasts should be created. The first three contrasts represent linear, quadratic and cubic trends in the data. [Any orthogonal trends could be used, but polynomial orthogonal trends are arbitrarily used here.] The final contrast employed is a sum vector that adds the responses of each participant over all four treatments. The orthogonal contrast matrix is presented in Table 9.

After the contrast matrix is generated, each contrast variable can be entered into the multiple regression equation in a hierarchical manner to determine the unique variance accounted for by each contrast. The results of the multiple regression analysis on the example data are presented in Table 10. Notice that the linear contrast variable (C_1) accounts for the same variance that is associated with the effects of the treatments (the interval component of the repeated measures ANOVA) in that the *SS* units and the R^2 values in the multiple regression analysis are exactly equal to the *SS* and eta^2 values in the summary tables for the repeated measures ANOVA presented in Table 3 and the classical ANOVA results presented in Table 4. The only other contrast variable to generate noteworthy consideration is the contrast variable C_4, as the *SS* and R^2 values for this contrast directly correspond to the *SS* and eta^2 values for the between subjects source of variation in the repeated measures ANOVA. Thus, the multiple regression approach generates results identical to the repeated measures ANOVA by utilizing a series of orthogonal contrast variables. Because all three of these methods (classical ANOVA, repeated measures ANOVA and multiple regression) generated the exact same *SS* partitions with the example data, the conceptual unity of the three approaches has been illustrated.

Multivariate Approach to Repeated Measures Analysis

The preceding portion of the paper was spent solely on examining the different univariate approaches to repeated measures analysis. A single factor repeated measures design, however, can be analyzed by using a multivariate approach. The multivariate approach, Multivariate Analysis of Variance (MANOVA), invokes a different sort of logic in completing the analysis. Rather than treating several measurements over time as a single dependent variable repeatedly measured, the multivariate approach treats the repeated measurements as separate dependent variables generated by one individual. Thus, in the example that has been used consistently throughout the paper, the multivariate approach would conceptually consider each of the four measures of depression as a separate dependent variable.

There are both advantages and disadvantages to conducting repeated measures analyses via the multivariate approach. One advantage of conducting this type of analysis is that the measurements are allowed to have any correlational structure (unlike in repeated measures ANOVA). That is, the sphericity assumption becomes unnecessary. This approach may more closely honor a given researcher's reality, provided that the researcher believes measurements to be correlated in a real world situation. In repeated measures ANOVA, the dependence among measures was considered to be of fixed form, and the researcher was penalized for analyzing data that did not exhibit equal correlations between measurements. In the multivariate approach, sphericity is not a consideration because each measurement is deemed a separate and unique dependent variable.

Table 9. Orthogonal Contrasts for Regression Analysis of Example Data in Table 2.

Subject	Dose	C_1	C_2	C_3	C_4	Y
1	0	-3	1	-1	43	1
2	0	-3	1	-1	43	3
3	0	-3	1	-1	42	3
4	0	-3	1	-1	41	4
5	0	-3	1	-1	45	5
1	150	-1	-1	3	43	10
2	150	-1	-1	3	43	6
3	150	-1	-1	3	42	8
4	150	-1	-1	3	41	8
5	150	-1	-1	3	45	9
1	300	1	-1	-3	43	14
2	300	1	-1	-3	43	15
3	300	1	-1	-3	42	11
4	300	1	-1	-3	41	13
5	300	1	-1	-3	45	13
1	450	3	1	1	43	18
2	450	3	1	1	43	19
3	450	3	1	1	42	20
4	450	3	1	1	41	16
5	450	3	1	1	45	18

Note. C_1 = contrast variable 1 (linear trend), C_2 = contrast variable 2 (quadratic trend), C_3 = contrast variable 3 (cubic trend), C_4 = sum vector, and Y = the dependent variable score.

Although discarding the sphericity assumption may sound enticing to even the most seasoned researcher, there are disadvantages to employing the multivariate approach to analyze repeated measures data. The primary disadvantage to utilizing the repeated measures approach is that statistical significance is difficult to obtain. Because the multivariate approach presumes that each measurement by an individual participant is a separate dependent variable, the advantage of repeatedly measuring participants is forfeited (i.e., a smaller number of participants necessary to generate results similar to classical ANOVA). Consequently, sample size issues become a paramount consideration, and a sample size that was sufficient to demonstrate an effect in a repeated measures ANOVA may be too small to demonstrate a similar effect using the multivariate approach.

The multivariate approach can be utilized in one of two ways: (a) by either transforming the score set into (k-1) difference variables and then analyzing the new variable set or (b) by creating a matrix of orthogonal or orthonormal coefficients, weighting each score by the corresponding coefficient, and analyzing the new matrix. Either method will generate identical results (Stevens, 1996), and there is no clear advantage to employing either method. The only stipulation in using either method is that only (k-1) new variables are created. The new matrix (either the differenced matrix or the transformed orthonormal matrix) is then analyzed using Hotelling's T^2 calculated by:

$$T^2 = n/(n\text{-}1)[(\mathbf{C}'\mathbf{M})'\,(\mathbf{C}'\textstyle\sum\mathbf{C})^{-1}\,(\mathbf{C}'\mathbf{M})]\ ,$$

where $\sum^{-1}$ is the inverse of the variance-covariance matrix, **M** is a column vector of means and **C** is a matrix of (k-1) orthogonal, orthonormal, or difference contrasts, **C**′ is the transpose of **C**. The calculated value of T^2 for the example data is 56.8181. This value can easily be transformed into an F_{calc} by invoking the formula:

$$F_{\text{calc}} = T^2\,(n - k + 1)/(k - 1)\ ,$$

where n is the number of participants and k is the number of treatment conditions. In this example, the

Table 10. Results of Multiple Regression Analysis with Table 2 Data.

Source	*SS*	*df*	*MS*	F_{calc}	*R*	R^2
C_1	625.00	1	625.00	94.70	.9730	.9467
C_2	0.00	1	0.00	0	0	0
C_3	0.00	1	0.00	0	0	0
Residual	35.20	18	1.96			
C_1	625.00	1	625.00	227.27	.9730	.9467
C_2	0.00	1	0.00	0.00	0	0
C_3	0.00	1	0.00	0.00	0	0
C_4	2.20	4	0.55	0.36	.0580	.0033
Residual	33.00	12	2.75			
Total	660.20					

computation is performed [$F_{\text{calc}} = 56.8181(5 - 4 + 1)/(4 - 1)$] and the transformed value of F_{calc} is equal to 37.879. The statistical significance of the F_{calc} value can be evaluated by using (k-1) and (n-k+1) degrees of freedom. For this example, the F_{crit} value is $F(3,2) = 99.164$ at the $\alpha = .01$ level. In contrast to the results of the univariate ANOVA, this result is not statistically significant primarily due to a radically smaller df_{RES} than in the univariate repeated measures ANOVA analysis.

Another reason statistical significance was not obtained by using the multivariate approach is that the sample size appears smaller. Because the ε value was 1.0 in the example data, the repeated measures ANOVA demonstrated the power equivalent of 20 participants (five participants measured on four occasions) whereas the multivariate approach presumed only five individuals measured on four occasions with four different dependent variables. The divergence in these two outcomes illustrates that the multivariate approach to repeated measures analysis is a more conservative method of analyzing effects and may not indicate statistically significant results even when other analytic methods do.

Are Univariate or Multivariate Methods Superior?

Repeated measures designs offer a number of advantages to social science researchers, one of which is the economy of research participants included in the study. As noted previously, the total size of the sample is decreased in the repeated measures case because the logic employed in the research design is different. That is, in a repeated measures design, participants serve as their own controls to eliminate the effects of individual participant error, thus necessitating a fewer number of participants. However, the advantages of the using the repeated measures design does not come without a cost. Disadvantages of this strategy include the introduction of carry-over, latency, and general practice effects, none of which can be fully removed from a given study (Keppel, 1991).

In regards to the analysis of repeated measures data, statisticians and researchers have debated whether univariate or multivariate approaches to repeated measures analyses are preferable with no clear consensus emerging from the discussions (Algina & Keselman, 1997; Barcikowski & Robey, 1984; Stevens, 1996). There are situations in which a univariate repeated measures ANOVA would be the most effective method of analyzing the data (e.g., if the sphericity assumption is satisfied and/or the between subjects and interval variance partitions account for the majority of the variance on the dependent variable) and other situations in which a multivariate approach would provide more favorable results (e.g., when the sphericity assumption is severely violated and/or when there is more variation in the treatment intervals than between the participants). Authors have recommended that both analyses be performed, because one or the other may be more powerful depending on the characteristics of the data (Barcikowski & Robey, 1984). This raises the question of which of the two analyses should be reported, and common sense indicates that if both analyses are completed, then this fact should be acknowledged to the consumers of the research. Most recently, however, the logic of employing both analytic strategies in the same analysis has been questioned (Keselman, Keselman, & Lix, 1995).

An important caveat must be noted. Multivariate methods were described earlier as offering an alternative to the univariate approach to analyzing repeated measures data and by which the sphericity assumption was not relevant. Although this comment is literally true, it is only accurate if the assumption of multivariate normality is met. Multivariate normality is the degree to which all linear combinations of several variables are normal (Henson, 1999). Further, simply ascertaining that variables are univariate and bivariate normal is not adequate to ensure that the system of variables is multivariate normally distributed. Rather, it is necessary to determine that each variable is normally distributed about fixed variables on all other variables.

Failing to assess multivariate normality can have deleterious consequences on statistical analyses, including those conducted with repeated measures designs. As noted by Marascuilo and Levin (1983, p. 203), "multivariate normality's impact and role . . . are basic to the inference procedures of multivariate analysis." Similarly, Thompson (1996b, p. 4) noted that, "Although multivariate normality is not required to estimate most multivariate parameters (e.g., function coefficients, structure coefficients), even in these cases the distributions of the variables must be reasonably comparable." Consequently, multivariate normality is an assumption that must be satisfied to ensure the accuracy and correct interpretation of multivariate results.

When multivariate normality has been met, however, the actual α level of the study is mathematically guaranteed to equal the preset α at the beginning of the study (Maxwell & Delaney, 1990). Consequently, if the multivariate normality assumption is not met, the α rate for the entire study will rarely equal the preset α level and may run as high as 10-20% (Tanguma, 1999). It is on this basis that Maxwell and Delaney recommended the use of the multivariate approach to repeated measures analysis if it is the researcher's desire is to avoid falsely rejecting the null hypothesis.

Higher-than-preset α rates are often a concern when the sphericity assumption is violated as well because many researchers do not adjust the degrees of freedom accordingly and completely overlook this very important statistical assumption. Even when sphericity is examined, there are only two options for adjusting for its violation: (a) adjust the degrees of freedom by a value of epsilon; or (b) use the multivariate approach. If the former is chosen as the method of choice, the results are only approximate because the epsilon value is an estimated correction factor.

Thus, the following general rules are posited based on the conjecture presented in previous research: (a) if sphericity is not violated, the univariate methods tend to be more powerful than the multivariate methods; (b) if sphericity is violated, neither method (univariate versus multivariate) tends to be preferable to the other, although a multivariate method can be used providing the variables are multivariate normal; and (c) if the size of the sample is substantially greater than the number of levels in the repeated variable, then the multivariate methods are preferred and tend to produce better results.

Summary and Conclusions

The present paper explored the analysis of univariate repeated measures designs by using both univariate and multivariate approaches. The univariate tests examined generated desirable results when certain statistical assumptions in the data were satisfied. The multivariate approach to repeated measures analysis was found to generate results generally free from the statistical assumptions in the univariate case but which tended to be a more conservative estimate of the effects. Both types of analyses were found to be valuable when certain situations were presented and when various assumptions were met.

If the assumption of sphericity is violated, the researcher is always able to employ the multivariate analysis as sphericity is not a required assumption in this approach. However, the multivariate method presumes that multivariate normality has been met, which is often not the case. The univariate approach is often more powerful than the multivariate approach when score variances are homogeneous as the df_{RES} is larger for the univariate test than for Hotelling's T^2. If the data are characterized by extreme variability, small effects of the treatment conditions may be hidden in the univariate data but elucidated by the multivariate approach. In situations where both the sphericity and normality assumptions are violated, multivariate rank-based procedures have been shown to control Type I errors and provide more statistical power than parametric procedures (Agresti & Pendergast, 1986; Beasley, in press). The bottom line, therefore, is that common analytic sense should guide practice, and researchers should critically examine their data to determine the most appropriate analytic strategy.

References

Agresti, A., & Pendergast, J. (1986) Comparing mean ranks for repeated measures data. *Communications in Statistics: Theory & Method*, *15*, 1417-1433.

Algina, J., & Keselman, H.J. (1997). Detecting repeated measures effects with univariate and multivariate statistics. *Psychological Methods, 2*, 208-218.

Barcikowski, R.S., & Robey, R.R. (1984). Decisions in single group repeated measures analysis: Statistical tests and three computer packages. *American Statistician, 38*, 148-150.

Beasley, T. M. (in press) Multivariate aligned rank test for interactions in multiple group repeated measures designs. *Multivariate Behavioral Research*.

Benton, R.L. (1991). Statistical power considerations in ANOVA. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 119-132). Greenwich, CT: JAI Press.

Box, G.E.P. (1954). Some theorems on quadratic forms in the study of analysis of variance problems: II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics, 25*, 484-498.

Edwards, A.L. (1979). *Multiple regression and the analysis of variance and covariance*. San Francisco: Freeman and Company.

Edwards, A.L. (1985). *Experimental design in psychological research* (5th ed.). New York: Harper & Row.

Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh, England: Oliver and Boyd.

Girden, E.R. (1992). *ANOVA: Repeated measures*. Newbury Park, CA: Sage.

Geisser, S., & Greenhouse, S.W. (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. *Annals of Mathematical Statistics, 29*, 885-891.

Greenhouse, S.W, & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95-112.

Haase, T., & Thompson, B. (1992, January). *The homogeneity of variance assumption in ANOVA: What it is and why it is required*. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX.

Henson, R.K. (1999). Multivariate normality: What is it and how is it assessed? In B. Thompson (Ed.), *Advances in social science methodology* (vol. 5, pp. 193-211). Stamford, CT: JAI Press.

Huynh, H., & Feldt, L.S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact *F*-distributions. *Journal of the American Statistical Association, 65*, 1582-1589.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Keppel, G., & Saufley, W.H. (1980). *Introduction to design and analysis: A student's handbook*. San Francisco: W.H. Freeman & Company.

Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs*. New York: W.H. Freeman & Company.

Keselman, H.J., Keselman, J.C., & Lix, L.M. (1995). The analysis of repeated measures measurements: Univariate tests, multivariate tests, or both? *British Journal of Mathematical and Statistical Psychology, 48*, 319-338.

Keselman, J.C., Lix, L.M., & Keselman, H.J. (1996). The analysis of repeated measurements: A quantitative research synthesis. *British Journal of Mathematical and Statistical Psychology, 49*, 275-298.

Marascuilo, L.A., & Levin, J.R. (1983). *Multivariate statistics in the social sciences: A researcher's guide*. Monterey, CA: Brooks/Cole.

Maxwell, S.E., & Delaney, H.D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

Rouanet, H., & Lepine, D. (1970). Comparison between treatments in a repeated-measurement design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology, 23*, 147-163.

Shavelson, R.J. (1988). *Statistical reasoning for the behavioral sciences* (2nd ed). Boston: Allyn & Bacon.

Snyder, P.A., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61*, 334-349.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Lawrence Erlbaum.

Tanguma, J. (1999). Analyzing repeated measures designs using univariate and multivariate methods. In B. Thompson (Ed.), *Advances in social science methodology* (vol. 5, pp. 233-250). Stamford, CT: JAI Press.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61*, 361-377.

Thompson, B. (1996a). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1996b). *Problems with multivariate normality: Can the multivariate bootstrap help*? Paper presented at the annual meeting of the Society for Applied Multivariate Research, Houston. (ERIC Document Reproduction Service No. ED 420 154)

Thompson, B. (1999a). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review, 11*, 157-169.

Thompson, B. (1999b). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory and Psychology, 9*, 167-183.

Wilcox, R.R. (1987). *New statistical procedures for the social sciences: Modern solutions to basic problems*. Hilldale, NJ: Lawrence Erlbaum.

Send correspondence to: Kevin M. Kieffer, Department of Psychology MC2127,
Saint Leo University, P.O. Box 6665, Saint Leo, FL 33574.
Email: kmkieffer@earthlink.net or Kevin.Kieffer@saintleo.edu.