

# MANOVA Method for Analyzing Repeated Measures Designs: An Extensive Primer

Ralph G. O'Brien

Statistics Department, University of Tennessee

Mary Kister Kaiser

Human Performance Center, University of  
Michigan

This article teaches the multivariate analysis of variance (MANOVA) method for repeated measures analysis to researchers who are already familiar with regular analysis of variance (ANOVA) methods. Repeated measures designs are traditionally analyzed with mixed-model ANOVAs. However, sphericity violations markedly affect the true Type I error rates and power for the mixed-model tests. However, tests based on the MANOVA approach are free of sphericity assumptions, and with modern computing software, the MANOVA approach is straightforward to use, even for complex designs and nontraditional hypotheses. We present a general strategy for implementing the MANOVA approach within statistical computing packages, and illustrate how regular (preplanned) tests and Scheffé-type simultaneous (post hoc) tests are conducted for a variety of designs and hypotheses.

Repeated measures designs involve each subject being measured  $p$  times on the same dependent variable. Repeated measurements may occur because each subject is measured at successive times (e.g., ages) or under several experimental conditions. Such conditions may result from the crossing or nesting of several factors called within-subjects or repeated factors. Factors composed of levels containing independent groups of subjects are called between-subjects or independent-groups factors. Repeated measures designs often have combinations of repeated and independent-group factors. Here we consider only intervally scaled dependent variables and procedures related to the analysis of variance (ANOVA). Because repeated observations are almost never independent (a critical assumption for regular ANOVA), repeated measures designs require special techniques for proper analysis. This article describes how multivariate anal-

ysis of variance (MANOVA) can be used to analyze such designs. We presume that you are familiar with ANOVA techniques for factorial designs, but not with MANOVA.

## Why Use the MANOVA Method

### *Problems With the Traditional Method*

In texts on univariate experimental design (e.g. Keppel, 1982; Kirk, 1982; Myers, 1979; Winer, 1971), presentations of repeated measures (or split-plot) designs focus on the so-called traditional univariate analysis. This method makes subjects a random-effects blocking factor, a step that leads to a mixed-model ANOVA. For review, consider a  $2 \times p$  design in which  $N$  Subjects (S) are nested within ( $w$ ) Gender (G), and every subject is measured at  $p$  Ages (A). The mixed-model test of the Gender main effect takes

$$F(G) = MS(G)/MS(SwG)$$

to be distributed as  $F[(p - 1), (N - 2)]$ . The test of the entire (omnibus) Age  $\times$  Gender interaction takes

$$F(A \times G) = MS(A \times G)/MS(A \times SwG) \quad (1)$$

to be distributed as  $F[(p - 1), (p - 1) \times (N - 2)]$ .

*Sphericity assumption.* If all of its distributional assumptions are met, the mixed-model approach is the most powerful method

---

Preparation of this article was supported by a Park National Bank faculty development fellowship to Ralph O'Brien. Mary Kaiser was supported by a National Science Foundation predoctoral fellowship and National Institute of Mental Health postdoctoral training grant MH-16892. We thank J. E. Keith Smith for a suggestion that resulted in Table 9, and two anonymous reviewers for suggesting several other improvements.

Requests for reprints should be sent to Ralph O'Brien, Department of Statistics, University of Tennessee, Knoxville, Tennessee 37996-0532.

for repeated measures analysis. Unfortunately, those assumptions are often troublesome. Specifically, the traditional analysis assumes sphericity (or circularity) among the repeated measurements: The variances and correlations of the measures are restricted in ways specific to the design and hypothesis being tested. (An explicit definition of sphericity involves more matrix algebra than we presume of our readers; see Rogan, Keselman, & Mendoza, 1979, sect. 4.) One form of sphericity is compound symmetry. This requires that all variances of the repeated measurements are equal, and that all correlations between the pairs of repeated measurements are equal. For example, if vocabulary size is measured at 2, 4, 6, and 8 years, compound symmetry requires the equality of the four population variances and the equality of the six unique correlations. Surely we would expect the correlations for adjacent ages (2:4, 4:6, and 6:8) to exceed the correlation between the measures taken at ages 2 and 8. Although violations of compound symmetry do not necessarily indicate violations of sphericity, they do imply that sphericity is unlikely. In general, sphericity is unnatural for most repeated measures data, and we believe that it is commonly violated in most designs with more than two repeated measurements.

*Effects from violating sphericity.* Regular (nonrepeated) balanced ANOVA is reasonably robust to violations of normality and equality of group variances. However, violations of the sphericity assumption often compromise the results of a mixed-model ANOVA for repeated measures. Box (1954) established that nonsphericity artificially inflates  $F$  values for omnibus tests of main effects and interactions involving the within-subjects factors. Therefore, the researcher who uses a mixed-model approach to analyze data that violate sphericity is testing these hypotheses using a Type I error rate that exceeds the stated alpha-level of the tests. This conclusion has been supported by many others; see Hearne, Clark, and Hatch (1983) and the references therein.

For example, Huynh and Feldt (1980) studied a design composed of one between-subjects factor with three groups and one within-subjects factor with five measurements. The variances of the five measurements were equal, but the correlation matrix was

$$\begin{bmatrix} 1.00 & & & & & \\ .80 & 1.00 & & & & \\ .60 & .80 & 1.00 & & & \\ .40 & .60 & .80 & 1.00 & & \\ .30 & .40 & .60 & .80 & 1.00 & \end{bmatrix}$$

Even if all the other assumptions are satisfied and the groups' sample sizes are infinitely large, the test of the Group  $\times$  Measurement interaction has a Type I error rate of .09 when the alpha level is presumed to be .05. Other cases result in even worse Type I error inflation.

These comments relate to the omnibus tests of the main effects and interactions. It is also important to understand the behavior of mixed-model tests for partial effects—comparisons, special contrasts, subeffects, and simple effects—because these are often the purest assessments of a researcher's hypotheses. Compared with the omnibus tests, tests of partial effects are even more unstable (Boik, 1981; Mitzel & Games, 1981). To help summarize these problems, let us return to the Gender  $\times$  Age design and consider the Linear Age Trend (Age-linear)  $\times$  Gender interaction. This subeffect of the Age  $\times$  Gender interaction can be tested using either a general error term or a specific error term. For the general error term,  $MS(A \times SwG)$ , one takes

$$F(A\text{-lin} \times G) = MS(A\text{-lin} \times G)/MS(A \times SwG)$$

to be distributed as  $F[1, (p - 1)(N - 2)]$ . For the specific error term,  $MS(A\text{-lin} \times SwG)$ , one takes

$$F(A\text{-lin} \times G) = MS(A\text{-lin} \times G)/MS(A\text{-lin} \times SwG)$$

to be distributed as  $F[1, (N - 2)]$ . The general error-term version assumes that sphericity holds for the omnibus test of  $A \times G$  and is quite unstable if that sphericity is violated. The specific error-term version has no sphericity assumption.

To elaborate, let us first consider the case when sphericity holds for the  $A \times G$  test. It can then be shown that  $MS(A \times SwG)$  and  $MS(A\text{-lin} \times SwG)$  have the same expected value. In fact,  $MS(A \times SwG)$  has the same

expected value as the specific error term for any subeffect of  $A \times G$ . The general error term differs from the specific terms in that  $MS(A \times SwG)$  is the average of the specific error terms for all possible one-degree-of-freedom subeffects of  $A \times G$ . If sphericity holds for the  $A \times G$  test, this average error term,  $MS(A \times SwG)$ , is a better estimate of error variance than are any of the individual specific error terms, such as  $MS(A\text{-lin} \times SwG)$ .  $MS(A \times SwG)$  has the most degrees of freedom; it uses the most information to estimate the error variance.

Once sphericity is violated, the optimality of the general error term vanishes. If sphericity is violated for the  $A \times G$  test,  $MS(A \times SwG)$  and  $MS(A\text{-lin} \times SwG)$  no longer have the same expected value.  $MS(A\text{-lin} \times SwG)$  maintains the proper expected value, regardless of the pattern of variances and correlations of the repeated measurements. However,  $MS(A \times SwG)$  becomes positively biased or negatively biased, depending on that pattern. As a result,  $F$  tests based on general error terms are out of control, producing either excessive Type I error rates or deficient power. To illustrate, Boik (1981) discussed two effects analogous to  $A\text{-lin} \times G$ , albeit with three groups instead of two. Using tests based on general error terms, the first effect had a .05-based error rate of .007; for the second it was .116. When moderately different means were specified, the test of the first effect showed a power of .25 versus .51 for the test based on the specific error term. Matters can get much worse, as demonstrated by Boik's figures displaying upper and lower bounds for Type I error rates. Even slight violations of sphericity can produce alarmingly unstable  $F$  tests. Boik summarized: "On the whole, the ordinary  $F$  tests have nothing to recommend them" (1981, p. 248). Like Mitzel and Games (1981) and Maxwell (1980), Boik recommended using specific error terms for each contrast. General error terms are average error terms, and they may be inappropriate for a given contrast or other partial effect.

How much power is sacrificed if a specific error term is used in a case where sphericity holds? Let us return to the Age-Linear  $\times$  Gender interaction assuming there are 15 males and 15 females. The critical values for this test are  $F[1, 84, .05] = 3.95$  using  $MS \times$

$(A \times SwG)$  as the error term, and  $F[1, 28, .05] = 4.20$  using  $MS(A\text{-lin} \times SwG)$ . Assuming a moderate effect size (specifically, a noncentrality of  $\lambda = 5.0$ ), then  $MS(A \times SwG)$  and  $MS(A\text{-lin} \times SwG)$  produce powers of .60 and .58, respectively. This power difference (.02) is a small price to pay for the stability gained from using the specific error term. In general, there is little power lost by the use of specific error terms, unless the sample sizes are quite small.

*Mauchly's test for sphericity.* Some authors suggested using a test for sphericity, Mauchly's  $W$ , to help choose between alternative repeated measures methods. However, this pretesting has several shortcomings. First, this test is quite sensitive to violations of normality, tending to accept sphericity too often for light-tailed distributions and reject sphericity too often for heavy-tailed distributions. Worse yet, these tendencies are amplified by increasing sample sizes (Huynh & Mandeville, 1979). Second, any acceptance of sphericity may simply be due to insufficient sample sizes. To use such an acceptance of sphericity to justify one's use of a mixed-model test is naive and careless. Third, some of the ANOVA effects may seem to satisfy their sphericity assumptions, whereas others may seem to violate them. Adopting different repeated measures methods for the various tests within a single design would produce a most confusing analysis. Finally, the concern and labor regarding sphericity and its pretest do not reap worthwhile increases in statistical power for the ANOVA effects. If all the conditions are right to produce a reliable and powerful pretest for sphericity, one may also assume that little power is sacrificed by immediately using ANOVA tests that do not assume sphericity. We agree with the assessment by Keselman, Rogan, Mendoza, and Breen (1980), who wrote about the "futility of attempting to assess the validity conditions" (p. 479) for the traditional model.

#### *Alternatives to the Traditional Model*

There are two parametric approaches that avoid sphericity assumptions for repeated measures analysis: (a) modifying the traditional mixed-model method, and (b) using multivariate analysis of variance methods. Neither approach is uniformly superior.

*Modifying the mixed-model F statistics.* Various modifications to the mixed-model solution have been investigated (Collier, Baker, Mandeville, & Hayes, 1967; Huynh, 1978; Huynh & Feldt, 1976; Wallenstein & Fleiss, 1979), and some perform quite well. Typically, these modifications involve systematically reducing the degrees of freedom for the sampling distribution of the  $F$  statistics: The amount of the reduction is dependent on the degree to which sphericity is violated. For example, consider  $F(A \times G)$  in Equation 1. Instead of using  $F[(p - 1), (p - 1)(N - 2)]$  as the sampling distribution, one takes  $F(A \times G)$  to be distributed as  $F[\epsilon(p - 1), \epsilon(p - 1)(N - 2)]$ , where  $\epsilon$  is bounded by  $1/(p - 1)$  and 1.0.  $\epsilon$  measures nonsphericity: If  $\epsilon = 1.0$  in the population, then sphericity holds, and the traditional sampling distribution is designated. Reductions in  $\epsilon$  indicate increasing degrees of nonsphericity and bring about suitable increases to the critical values for  $F$ . Several estimators of  $\epsilon$  have been developed, particularly by Huynh. Unfortunately, many researchers may find this work too young to use effectively. Articles on these methods are mostly quite technical, and they are primarily concerned with omnibus tests in simpler types of factorial designs. Although it is possible to generalize these methods to handle complicated designs and special partial effects, those generalizations are too difficult for most researchers to handle on their own. In addition, these advances are just beginning to be incorporated into the popular statistical computing packages: BMDP2V now uses Huynh and Feldt's (1976) best estimator,  $\tilde{\epsilon}$ . An updated version of SPSS (Statistical Package for the Social Sciences)-MANOVA also incorporates some modern  $\epsilon$ -adjustments (Burns, 1984).

*Using MANOVA.* The MANOVA approach does not suffer from either the generalization gap or the software gap. Once you understand the basic logic, you should be able to apply it to any design and hypothesis test, using one of the many MANOVA routines that are already available within the major statistical packages. The MANOVA approach is most flexible and straightforward. There are no questions regarding proper error terms or modifications to the degrees of freedom. However, its greatest virtue is the following:

In the strictest sense, the MANOVA approach is the natural generalization of the use of the specific type of error term for contrasts with 1 degree of freedom. Of course, any ANOVA effect can be defined by a suitable set of contrasts taken over all the cells of the design. The MANOVA approach handles sets of contrasts in such a way that each contrast in the set remains linked with just its specific error term. As a result, we avoid all the problems associated with general (average) error terms.

Several studies have compared the power of the modified traditional tests and the MANOVA approach (Davidson, 1972; Huynh, 1978; Mendoza, Toothaker, & Nicewander, 1974; Rogan et al., 1979). Even though only a limited number of situations have been investigated, this work found that no procedure is uniformly the most powerful or even usually the most powerful. Besides being related to the groups' sample sizes, the power differences depend on technical relations among many factors that researchers have little knowledge of, such as the pattern of population means, the variances and correlations of the repeated measures, and the real distribution of the measures (which we pretend is multivariate normal). Because no clear-cut power differences exist, the best strategy is to choose a single method and seek to master it. We believe that method should be the MANOVA approach.

#### How to do Repeated Measures Analysis With MANOVA

This article now discusses how to carry out a MANOVA-based repeated measures analysis. Much of this material can be found in other sources, such as Timm (1980a, 1980b), but our presentation is much less technical. To learn to do repeated measures analysis, you need not plow through ground congested with long matrix equations, non-negative definite quadratic forms, eigenvalues, and the union-intersection principle. Even though we avoid such matters, our discussion of repeated measures per se is more extensive than what is available in applied multivariate analysis texts. Those books typically treat repeated measures as a side issue and only discuss simple designs and standard hypotheses—preplanned main effects and interactions. We discuss designs ranging from

simple to moderately complex, and we demonstrate many types of hypothesis tests, including contrasts, subeffects, and simple effects, as well as procedures for familywise error protection. In many ways, the level and style of our instruction resembles that given by McCall and Appelbaum (1973), who concentrated on preplanned tests of polynomial trends.

To keep technical matters down to earth, we include programming statements for one popular statistical computing package, SPSSX, and its MANOVA procedure. SPSSX is widely available, and its syntax is close enough to competing software (such as Statistical Analysis System, or SAS, and its Procedure General Linear Model, or PROC GLM, routine) that experienced users of other packages should have little difficulty making the necessary translations. To aid such translations, we avoid using the SPSSX-MANOVA commands that specifically deal with repeated measures analysis. Although these commands are excellent tools for those who already understand MANOVA-based repeated measures analysis, they let novices skip steps important in learning the simplicity and generality of these methods.

*Design With Two Groups and Two Measurements*

Let us first consider the design in which subjects are randomly assigned to one of two groups (control and experimental) and are measured twice: before and after the experimental group first received the treatment. Table 1 illustrates this design and exemplifies how we display all repeated measures designs. The rows identify the independent groups defined by the between-subjects factors. The columns display the repeated measurements defined by the within-subjects factors. This design is often analyzed using a two-group analysis of covariance (ANCOVA) model with the pretest as the covariate and the posttest as the dependent variable. Bock (1975, pp. 489-496) discussed how the ANCOVA approach differs from the repeated measures approach. Only repeated measures approaches are dealt with here.

*General two-phase strategy applied to the interaction test.* This design usually requires a test of the Group  $\times$  Time interaction: Is

Table 1  
*Two Groups and Two Measurements*

Subject	Time	
	Pretest ( $y_{r1}$ )	Posttest ( $y_{r2}$ )
Control		
1	2	3
2	4	3
3	6	5
4	5	3
5	4	6
Experimental		
6	8	9
7	5	8
8	3	5
9	4	4

the mean change from pretest to posttest the same for the two groups? Letting  $\mu_{ij}$  be the population mean for the  $i$ th group and  $j$ th measurement, this hypothesis is

$$H_0: (\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22}) = 0.$$

How can we test it?

First, consider each subject to be a unique, single-subject experiment. Let  $y_{r1}$  and  $y_{r2}$  be the pretest and posttest scores for Case  $r$ , and let  $d_r = (y_{r2} - y_{r1})$  be the amount of change. Consider Subject 3 from the control group and Subject 7 from the experimental group:

$$y_{31} = 6 \quad y_{32} = 5, \quad \text{thus } d_3 = -1$$

$$y_{71} = 5 \quad y_{72} = 8, \quad \text{thus } d_7 = +3.$$

It appears that the treatment could be increasing the scores, but such an inference is unreasonable with just one case from each group.

Obviously, we need to study the amount of change over all the subjects. The sample sizes, average amounts of change, and variances for the amount of change for the two groups are

$$n_1 = 5 \quad \bar{d}_1 = -0.20 \quad s_1^2(d) = 2.70$$

$$n_2 = 4 \quad \bar{d}_2 = 1.50 \quad s_2^2(d) = 1.67.$$

Now we can ask a more familiar question: Is  $\bar{d}_1 = -.20$  significantly different from  $\bar{d}_2 = 1.50$ ? Presuming that the  $d$  variable satisfies the necessary assumptions, a two-group  $t$  test

can be used, or, equivalently, we can use the  $F$  test from a two-group ANOVA. Using these data, we get

$$F(1, 7) = \text{MSBG}/\text{MSWG} \\ = 6.42/2.26 = 2.85; \quad p = .14.$$

There is no strong evidence that the mean amount of change over time differed between the two groups: the Time  $\times$  Group interaction is nonsignificant.

This test is accomplished in SPSSX by first entering the values for GROUP (coded 1 and 2), PRETEST, and POSTTEST; then computing the amount of change for each subject with COMPUTE TIMEDIFF = POSTTEST - PRETEST and finally performing the test for group differences with

```
MANOVA TIMEDIFF BY GROUP (1, 2)/
PRINT = CELLINFO(MEANS)/
METHOD = SSTYPE(UNIQUE)/
DESIGN = CONSTANT, GROUP/.
```

This testing of the Time  $\times$  Group interaction illustrates simply the two distinct phases: the within-subjects phase and the between-subjects phase. The within-subjects phase is: Disregarding the between-subjects factors, compute the contrast variable(s) for the within-subjects part of the effect of interest. Here we only need one contrast variable, TIMEDIFF. Later we need multiple contrast variables to compare more than two measurements. The between-subjects phase is: Use the contrast variable(s) as the dependent variables in a regular (nonrepeated) univariate (or multivariate) analysis of variance. The design for this analysis includes just the between-subjects factors. (The contrast variables handle the within-subject factors.) Compute test statistics for the between-subjects part of the effect of interest. Here the between-subjects test is the group effect. Using the time contrast variable and testing the group effect gives us the Time  $\times$  Group interaction.

*Testing the main effects.* Let us also outline the tests for the Group and Time main effects, even though they are usually irrelevant for this particular design. For the Group main effect, the within-subjects part requires

us to construct a contrast variable that averages over time:

$$\text{COMPUTE AVERAGE} \\ = (\text{PRETEST} + \text{POSTTEST})/2.$$

The between-subjects part is again the Group effect. Testing for group differences on the AVERAGE dependent variable leads to a test of the Group main effect.

For the Time main effect, the within-subjects part is the Time effect, so the contrast variable needs to be TIMEDIFF again. The between-subjects part requires us to average over the groups (i.e., to test the grand mean). In SPSSX-MANOVA, this is the test for CONSTANT, which occurred previously in the DESIGN = CONSTANT, GROUP/ statement. For unbalanced designs, CONSTANT can be defined in several ways. Using the command METHOD = SSTYPE(UNIQUE)/ weights the groups equally and produces a test of the unweighted hypothesis:

$$H_0: (\mu_{11} - \mu_{12})/2 + (\mu_{21} - \mu_{22})/2 = 0.$$

Without the UNIQUE specification, the CONSTANT test corresponds to the weighted hypothesis:

$$H_0: 5(\mu_{11} - \mu_{12})/9 + 4(\mu_{21} - \mu_{22})/9 = 0.$$

In the remaining examples, we will make few comments regarding the analysis of unbalanced designs. As long as every subject has data for all repeated measurements, the within-subjects phase is unaffected. For the between-subjects phase, the issues regarding unbalanced group sizes are not different from those concerning nonrepeated fixed-effects designs. (See Herr & Gaebelein, 1978, for a cogent summary of those issues.) When a design involves group sizes that are unrelated to the hypotheses being addressed, we generally prefer to test unweighted hypotheses, which we do throughout the rest of the article. In your own work, you may have some cause to do otherwise. Nevertheless, the general two-phase strategy still applies.

*Testing simple effects.* At this point you should not be surprised to see how the two-phase strategy applies to simple effects. First consider the Group within Posttest simple effect:

$$H_0: \mu_{12} - \mu_{22} = 0.$$

Table 2  
Summary of Tests for Two Groups and Two Measurements

Between-groups test	Dependent variable used			
	AVERAGE	TIMEDIFF	PRETEST	POSTTEST
CONSTANT (grand mean = 0) GROUP (group main effect)	Group main effect	Time main effect Group × Time	Group within pretest	Group within posttest
CONSPLUS GROUP (1) (control mean = 0)		Time w/in control group		
CONSPLUS GROUP (2) (experimental mean = 0)		Time w/in exper. group		

Here the within-subjects part requires the use of POSTTEST as the dependent variable, and the between-subjects part calls for a test of the Group effect. Similarly, if we use PRETEST as the dependent variable in the two-group ANOVA, we get the test of Group within Pretest.

Another simple effect is Time within Experimental Group:

$$H_0: \mu_{21} - \mu_{22} = 0.$$

Clearly, the appropriate contrast variable is TIMEDIFF. We must test whether the experimental group has a mean for TIMEDIFF that is different from zero. If we are satisfied that the variance of TIMEDIFF is the same in both groups, we should use a test based on the pooled variance, estimated by the within-group mean square (MSWG), giving

$$F(1, 7) = 4(1.5)^2/2.26 = 3.99; p = .086.$$

The MANOVA statements to execute this test are:

```
MANOVA TIMEDIFF BY GROUP (1, 2)/
METHOD = SSTYPE(UNIQUE)/
DESIGN = CONSPLUS GROUP(1),
CONSPLUS GROUP(2)/.
```

CONSPLUS GROUP(2) is the Time within Experimental Group simple effect; CONSPLUS GROUP(1) is the Time within Control Group simple effect. If the group variances for TIMEDIFF are not the same, then we should use a test based on the separate variances. For example the Time within Experimental Group test would yield  $F(1, 3) = 4(1.5)^2/1.67 = 5.39; p = .10$ .

*Conclusion.* Table 2 summarizes all the tests discussed for the two-group, two-measurement design. Every test for this design is univariate because each hypothesis has a within-subjects part with just 1 degree of freedom. It is worth mentioning that whenever a hypothesis has this characteristic, the two-phase strategy produces the same univariate test as the mixed-model strategy does using specific error terms. We now turn to a design that requires multivariate analyses.

*Design With Three Groups and Three Measurements*

We now extend the previously discussed design by adding a third group and measurement. The groups are now (1) Control, (2) Treatment A, and (3) Treatment B; and each subject has measures for Pretest ( $y_1$ ), Posttest ( $y_2$ ), and Follow-up ( $y_3$ ). Table 3, an expansion of Table 1, illustrates this design.

*Vector algebra.* It will help (especially for more complicated designs) to use a little vector algebra. Let the vectors  $y$ , and its transpose,  $y'$ , represent the set of repeated measurements:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \quad y' = [y_1 \ y_2 \ y_3].$$

All of the contrast variables that we use to handle the within-subjects parts of the various hypotheses can be represented in terms of simple vector multiplication:  $[a_1 \ a_2 \ a_3]y = a_1y_1 + a_2y_2 + a_3y_3$ . For example,  $[-1 \ +1 \ 0]y = -y_1 + y_2$  expresses the con-

Table 3  
Three Groups and Three Measurements

Subject	Time		
	Pretest	Posttest	Follow-up
Control			
1	2	3	3
2	4	3	4
3	6	5	7
4	5	3	4
5	4	6	4
Treatment A			
6	8	9	9
7	5	8	9
8	3	5	6
9	4	4	5
Treatment B			
10	4	7	8
11	3	5	6
12	6	9	8
13	6	6	8
14	2	5	6
15	3	7	7
16	5	7	8

trast variable that compares the Posttest with the Pretest.

Applying MANOVA to the test of interaction. As in the 2 × 2 design, the most important omnibus test for this 3 × 3 design is likely to be the Time × Group interaction. To perform this test with the two-phase strategy, we need first to ask: What is the within-subjects part of the Time × Group interaction? Of course, it is the Time effect. Because Time has three levels, the Time effect has 2 degrees of freedom and therefore requires two contrast variables; for example:

$$d_1 = y_2 - y_1 = [-1 \ 1 \ 0]y$$

$$d_2 = y_3 - y_1 = [-1 \ 0 \ 1]y.$$

The between-subjects part of this hypothesis is the Group effect, which also has 2 degrees of freedom. Putting the two parts together generates the test for the Time × Group interaction: Do the three groups have different mean changes over time?

The SPSSX statements are still straightforward. The within-subjects part becomes

COMPUTE PSTVSPRE

= POSTTEST - PRETEST

COMPUTE FOLVSPRE

= FOLLOWUP - PRETEST

The between-subjects phase uses

MANOVA PSTVSPRE, FOLVSPRE

BY GROUP(1, 3)/

METHOD = SSTYPE(UNIQUE)/

DESIGN = CONSTANT, GROUP/

Examine Tables 4 and 5 to see the key portions of the output produced by these commands. There are four alternative multivariate test statistics: Pillai's, Hotelling's, Wilks's lambda, and Roy's. Pillai's statistic, sometimes called the Pillai-Bartlett, was first proposed by Nanda (1950). Olson (1974, 1976) compared these statistics, along with other statistics, and recommended Pillai's for general use. However, none of these statistics is uniformly better than the others. All have the same distributional assumptions, but each has a unique way of obtaining univariate test statistics for multivariate hypotheses. This gives them somewhat different characteristics in terms of power and sensitivity to violations of assumptions. Fortunately, however, they usually tell the same story about one's data.

Pillai's, Hotelling's, and Wilks's lambda statistics all have similar roots. Most important, they all have good methods to convert them to approximate *F* statistics (see Srivas-

Table 4  
Time × Group Test for the 3 × 3 Design  
(SPSSX-MANOVA Output)

Test name	Value	<i>F</i> <sup>a</sup>	Hypothesis <i>df</i>	Error <i>df</i>
Pillai	0.643	3.08*	4.0	26.0
Hotelling	1.782	4.90**	4.0	22.0
Wilks	0.359	4.02***	4.0	24.0
Roy	0.640			

Note. Dependent variables: PSTVSPRE FOLVSPRE. Effect: Group. Multivariate tests: (*s* = 2, *M* = -.5, *N* = 5).

<sup>a</sup> Approximate.

\* *p* = .033.

\*\* *p* = .006.

\*\*\* *p* = .012.



Table 5  
Univariate *F* Tests for the 3 × 3 Design  
(SPSSX-MANOVA output)

Dependent variable	Hypothesis SS	Error SS	Hypothesis MS	Error MS	<i>F</i>
PSTVSPRE	20.24	25.51	10.12	1.96	5.16*
FOLVSPRE	25.59	14.41	12.80	1.11	11.55**

Note. Univariate *F* tests with 2, 13 *df*. *SS* = sum of squares. *MS* = mean square.

\* *p* = .022.

\*\* *p* = .001.

tava & Khatri, 1979), making them seem familiar to most researchers. Anyone familiar with univariate ANOVA can examine Table 4 and see that all three *F* statistics are significant at the .05 level. In reporting such statistics, researchers hardly need to broach the fact that a multivariate analysis was done. Note especially that all three *F* statistics have 4 degrees of freedom for their hypothesis—just what you should expect for the interaction test for a 3 × 3 design. In fact, the hypothesis degrees of freedom makes sense for all multivariate tests of repeated measures designs. In contrast, also note in Table 4 that the these *F* approximations have different values for their error degrees of freedom; sometimes they can even be noninteger.

Roy's statistic is often called the greatest characteristic root statistic or the largest root statistic. No general way exists yet to convert it to an *F* statistic. Instead, one must use tables or charts to obtain its critical values. Because these values are important in performing Scheffé-type tests, we need to discuss how to find them. The three parameters for degrees of freedom for Roy's statistic—*s*, *m*, and *n*—are based on four aspects of the test of interest: (a) *Wdf*: degrees of freedom for the within-subjects part of the effect, in other words the number of contrast (dependent) variables being used for the test; (b) *Bdf*: degrees of freedom for the between-subjects part of the test; (c) *g*: the number of independent groups of subjects; and (d) *N*: the total sample size.

Now we can define Roy's parameters: *s* = the minimum of *Wdf* and *Bdf*; *m* = (|*Wdf* - *Bdf*| - 1)/2; and *n* = (*N* - *g* - *Wdf* - 1)/2. The Time × Group interaction test has *Wdf* =

2, *Bdf* = 2, *g* = 3, and *N* = 16. Thus, we get *s* = 2, *m* = -.5, and *n* = 5, as Table 4 shows. From Harris's (1975) Table A5, the critical values for the .05 level and the .01 level are  $\theta[2, -.5, 5, .05] = .498$  and  $\theta[2, -.5, 5, .01] = .623$ . The sample value,  $\theta_1 = .640$ , exceeds the .01 critical value.

A word of caution: Some statistical packages report Roy's statistic in a different form. For example, the MANOVA option in SAS PROC GLM defines  $\lambda_1 = \theta_1/(1 - \theta_1)$  to be Roy's maximum root. To conform with our presentation, simply compute  $\theta_1 = \lambda_1/(1 + \lambda_1)$ . One can also compute  $\theta_1$  by taking the square root of the canonical correlation reported with the other multivariate results.

When *s* = 1, Pillai's, Hotelling's, Wilks's lambda, and Roy's statistics are functionally identical and convert to the same *F* statistic, which has an exact *F* distribution. The *F* transform for Wilks's lambda is also exact for *s* = 2.

*Testing the main effects.* The tests of the group and time main effects follow directly from their counterparts in the 2 × 2 design and are summarized in Table 6. (We included the raw data in Table 3 and the many *F* statistics in Table 6 so that you can practice these methods by replicating some of our analyses.) The time main effect requires a multivariate test.

*Testing subeffects.* The univariate tests shown in Table 5 are tests of subeffects of the Group × Time interaction. The test using only PSTVSPRE examines whether the three Groups average the same change from the Pretest to the Posttest—the Group × Time [-1 1 0] subeffect. The test using FOLVSPRE is the Group × Time [-1 0 1] subeffect. Because these are univariate tests, they are equivalent to those obtainable from a mixed-model approach using specific error terms.

This is a good place to answer a question often asked by nonstatisticians. The Group × Time interaction analysis contrasted the Posttest with the Pretest (PSTVSPRE), and the Follow-up with the Pretest (FOLVSPRE), but we did not contrast the Follow-up with the Posttest by using FOLVSPST = FOLLOWUP - POSTTEST as a third contrast variable. Are we losing something? No. To see that FOLVSPST cannot be useful as a third dependent variable, note that FOLVSPST =

Table 6  
 Summary of Various Tests for Three Groups and Three Measures

Between-groups test	Dependent variable(s) used			
	Univariate AVERAGE [1 1 1]y	Multivariate PSTVSPRE & FOLVSPRE [-1 1 0]y & [-1 0 1]y	Univariate PSTVSPRE [-1 1 0]y	Univariate FOLVSPRE [-1 0 1]y
CONSTANT (grand mean = 0)		Time main effect, $F(2, 12) = 22.69$	Time[-1 1 0] subeffect, $F(1,13) = 11.95$	Time[-1 0 1] subeffect, $F(1,13) = 47.61$
GROUP (group main effect)	Group main effect, $F(2, 13) = 2.91$	Group $\times$ Time interaction, $F(4, 26) = 3.08$	Group $\times$ Time [-1 1 0], $F(2, 13) = 5.16$	Group $\times$ Time [-1 0 1], $F(2, 13) = 11.55$
GROUP (1) (group [-2 1 1] subeffect)	Group[-2 1 1] subeffect, $F(1, 13) = 5.82$	Group[-2 1 1] $\times$ Time interaction subeffect, $F(2, 12) = 8.71$	Group[-2 1 1] $\times$ Time[-1 1 0], $F(1, 13) = 8.00$	Group[-2 1 1] $\times$ Time[-1 0 1], $F(1, 13) = 18.86$
GROUP (2) (group [0 -1 1] subeffect)	Group[0 -1 1] subeffect, $F(1, 13) = 0.07$	Group[0 -1 1] $\times$ Time interaction subeffect, $F(2, 12) = 0.87$	Group[0 -1 1] $\times$ Time[-1 1 0], $F(1, 13) = 1.12$	Group[0 -1 1] $\times$ Time[-1 0 1], $F(1, 13) = 1.83$

Note. Multivariate tests are based on Pillai's statistic.

FOLVSPRE - PSTVSPRE. Therefore, if the three groups have the same population means for both FOLVSPRE and PSTVSPRE, they must also have the same means for FOLVSPST. The FOLVSPST contrast is already contained in the multivariate combination of FOLVSPRE and PSTVSPRE, making FOLVSPST redundant. In fact, any third contrast variable would be redundant. For this Group  $\times$  Time test, any pair of (nonredundant) contrast variables for the time effect will produce the same multivariate test statistics. It makes no difference whether we use PSTVSPRE and FOLVSPRE, PSTVSPRE and FOLVSPST, FOLVSPRE and FOLVSPST, or even some other pair such as

$$\begin{aligned} \text{PREVSAFT} &= 2 * \text{PRETEST} - \text{POSTTEST} \\ &\quad - \text{FOLLOW-UP} \\ \text{FOLVSPST} &= \text{FOLLOW-UP} - \text{POSTTEST}. \end{aligned}$$

With a little forethought, you can create contrast variables that will give some of the specific univariate tests important for your particular analysis. If you need to perform even more such tests, you must run them in groups of no more than *Wdf* nonredundant contrast variables per group.

These interaction subeffects can be made more specific by forming contrasts across the

three groups. For instance, let Group [-2 1 1] represent the contrast that compares the control group's mean with the unweighted average of the treatment groups' means. Here, *means* denotes the means for the time-effect contrast variables. In SPSSX-MANOVA, this contrast is defined and tested as GROUP (1) in the following commands:

```
MANOVA  PSTVSPRE, FOLVSPRE
                BY GROUP (1,3)/
METHOD = SSTYPE(UNIQUE)/
PARTITION(GROUP) = (1,1)/
CONTRAST(GROUP) = SPECIAL (1 1 1,
                -2 1 1, 0 -1 1)/
DESIGN = CONSTANT = 0, GROUP(1),
                GROUP(2)/.
```

GROUP (2) is the Group [0 -1 1] contrast. Together, GROUP (1) and GROUP (2) saturate the 2 degrees of freedom for the Group factor.

Interpreting the output again involves the pairing of the contrast variables with the between-groups test. Table 6 summarizes all tests produced by the previously discussed SPSSX-MANOVA commands. The multivariate

test of the GROUP(1) contrast with both PSTVSPRE and FOLVSPRE effectively tests the Group[-2 1 1] × Time interaction subeffect. It examines whether the control group's time profile is the same as the treatment groups' average time profile. The univariate test of the GROUP(1) contrast with just FOLVSPRE tests a most specific interaction subeffect: Group[-2 1 1] × Time[-1 0 1].

You should also now see how to construct subeffects for the Group and Time main effects. Just use the two-phase strategy of constructing the appropriate dependent contrast variables and pairing them with the between-groups test of interest. Table 6 gives examples.

*Simple effects.* General simple effects can be tested in much the same manner as we did for the 2 × 2 design. However, now we can also quickly carry out subeffects of the simple effects. For example, the test of Group [-2 1 1] within Follow-up can be performed by using just FOLLOW-UP as the dependent variable and testing the Group[-2 1 1] contrast as previously shown. The test of Time [-1 1 0] within Treatment B can be performed by testing whether the mean for PSTVSPRE differs from zero for the third group: The issues and mechanics regarding this test extend readily from those discussed in the 2 × 2 design.

*Design With Two Between-Subjects Factors and Two Within-Subjects Factors*

We now enlarge the 3 × 3 design by adding a second between-groups factor, Gender (1 = Male, 2 = Female), and a second within-subjects factor, Hour of measurement (Hour = 1, 2, 3, 4, or 5), created because each case is measured five times during a single testing session. These five times are equally spaced. Factors like Hour are found in studies involving measurements that change systematically over a short period of time, such as subjects' reaction times following their initial intake of alcohol. To avoid confusion, the within-subjects factor composed of Pretest, Posttest, and Follow-up is now called *Phase*. The between-subjects factor composed of Control, Treatment A, and Treatment B is now called *Treatment*. Table 7 shows the design, which is an expansion of Table 3.

Table 3 is reproducible from Table 7 if one averages over the 5 hr and ignores the gender factor.

The analysis of this design proceeds in the same manner as for the 3 × 3 design, albeit with some added complexities. First, we must define appropriate sets of contrast variables to handle both within-subjects effects. Sets of variables must be defined for the Phase effect, the Hour effect, the Phase × Hour interaction, and the average over Phase and Hour. To develop these four sets of contrast variables methodically, it will help to use another kind of vector multiplication.

*Direct product of two vectors.* The following operation is called a direct product (or Kronecker product):  $[a_1 a_2] \otimes [b_1 b_2 b_3] = [a_1b_1 a_1b_2 a_1b_3 a_2b_1 a_2b_2 a_2b_3]$ . For example:

$$\begin{aligned} [2 \ 7] \otimes [1 \ 6 \ 3 \ 9] \\ &= [2(1) \ 2(6) \ 2(3) \ 2(9) \ 7(1) \ 7(6) \ 7(3) \ 7(9)] \\ &= [2 \ 12 \ 6 \ 18 \ 7 \ 42 \ 21 \ 63]. \end{aligned}$$

If the first vector has  $e_1$  elements and the second has  $e_2$  elements, then the direct product is a vector with  $e_1e_2$  elements. When forming the successive products, the  $b$  elements cycle most rapidly.

*Constructing contrast variables with direct products.* Let  $y_{jk}$  be the measure taken during Phase  $j$  at Hour  $k$  and let the 15 elements of the observation vector be arranged in the natural order:

$$y' = [y_{11} \ y_{12} \ y_{13} \ y_{14} \ y_{15} \ y_{21} \ \dots \ y_{35}],$$

so that the right-most subscript ( $k$ , for hour) is cycling fastest. Every contrast variable that we construct will have the form

$$d = [c_{11} \ c_{12} \ c_{13} \ c_{14} \ c_{15} \ c_{21} \ \dots \ c_{35}]y,$$

where the contrast coefficients,  $c_{jk}$ , come from the direct product of a contrast over the phase levels × a contrast over the hour levels:

$$\begin{aligned} [\text{Contrast for phase}] \otimes [\text{contrast for hour}] \\ &= [p_1 \ p_2 \ p_3] \otimes [h_1 \ h_2 \ h_3 \ h_4 \ h_5] \\ &= [p_1h_1 \ p_1h_2 \ p_1h_3 \ p_1h_4 \ p_1h_5 \ p_2h_1 \ \dots \ p_3h_5]. \end{aligned}$$

We will do the phase effect variables first. As in the 3 × 3 design, we can compare the three phases using the contrast vectors [-1 1 0] and [-1 0 1]. For the Phase effect,

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

we need to average across the five levels of Hour, so we use [.2 .2 .2 .2 .2] which is functionally equal to [1 1 1 1 1]. Pairing the two phase-effect contrasts with the average-over-hour contrast yields the two contrast variables for the Phase effect:

$$P1 = [-1 1 0] \otimes [1 1 1 1 1]y$$

$$= [-1 -1 -1 -1 -1 1 1 1 1 1 0 0 0 0 0]y$$

$$P2 = [-1 0 1] \otimes [1 1 1 1 1]y$$

$$= [-1 -1 -1 -1 -1 0 0 0 0 0 1 1 1 1 1]y.$$

In SPSSX, P1 is constructed using COMPUTE  $P1 = -(Y11 + Y12 + Y13 + Y14 + Y15) + (Y21 + Y22 + Y23 + Y24 + Y25)$ .

The contrast variables for the Hour effect are constructed in a similar way. Because the Hour effect has quantitative levels, it makes a natural candidate for orthogonal polynomial

trend contrasts. Using [1 1 1] to average over the Phase factor, we form

$$H1 = [1 1 1] \otimes [-2 -1 0 1 2]y$$

$$= [-2 -1 0 1 2 -2 -1 0 1 2 -2 -1 0 1 2]y$$

$$H2 = [1 1 1] \otimes [2 -1 -2 -1 2]y$$

$$H3 = [1 1 1] \otimes [-1 2 0 -2 -1]y$$

$$H4 = [1 1 1] \otimes [1 -4 6 -4 1]y.$$

H1 is the contrast variable for Hour (linear), H2 constructs Hour (quadratic), H3 constructs Hour (cubic), and H4 constructs Hour (quartic). The coefficients for polynomial trend contrasts can be found in many sources, including Keppel (1982), Kirk (1982), and Winer (1971).

Table 7  
3 × 2 Groups and 3 × 5 Measurements

Treatment × Gender	Phase × Hour														
	Pretest					Posttest					Follow-up				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Control															
Male subjects															
1	1	2	4	2	1	3	2	5	3	2	2	3	2	4	4
2	4	4	5	3	4	2	2	3	5	3	4	5	6	4	1
3	5	6	5	7	7	4	5	7	5	4	7	6	9	7	6
Female subjects															
4	5	4	7	5	4	2	2	3	5	3	4	4	5	3	4
5	3	4	6	4	3	6	7	8	6	3	4	3	6	4	3
Treatment A															
Male subjects															
6	7	8	7	9	9	9	9	10	8	9	9	10	11	9	6
7	5	5	6	4	5	7	7	8	10	8	8	9	11	9	8
Female subjects															
8	2	3	5	3	2	2	4	8	6	5	6	6	7	5	6
9	3	3	4	6	4	4	5	6	4	1	5	4	7	5	4
Treatment B															
Male subjects															
10	4	4	5	3	4	6	7	6	8	8	8	8	9	7	8
11	3	3	4	2	3	5	4	7	5	4	5	6	8	6	5
12	6	7	8	6	3	9	10	11	9	6	8	7	10	8	7
Female subjects															
13	5	5	6	8	6	4	6	6	8	6	7	7	8	10	8
14	2	2	3	1	2	5	6	7	5	2	6	7	8	6	3
15	2	2	3	4	4	6	6	7	9	7	7	7	8	6	7
16	4	5	7	5	4	7	7	8	6	7	7	8	10	8	7

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 8  
Summary of Various Tests for 3 × 2 Groups and 3 × 5 Measures

Between-groups test	Dependent variables used				
	AVERAGE	P1 & P2	H1 H2 H3 H4	P1 × H1 . . . P2 × H4	P1 × H1 . . . P2 × H2
CONSTANT		P $F(2, 9) = 19.6$	H $F(4, 7) = 24.3$	P × H $F(8, 3) = 0.48$	P × H[ln + qd] $F(4, 7) = 1.16$
T	T $F(2, 10) = 3.94$	T × P $F(4, 20) = 2.67$	T × H $F(8, 16) = 0.38$	T × P × H $F(16, 8) = 0.25$	T × P × H[ln + qd] $F(8, 16) = 0.39$
G	G $F(1, 10) = 3.66$	G × P $F(2, 9) = 0.32$	G × H $F(4, 7) = 0.90$	G × P × H $F(8, 3) = 0.92$	G × P × H[ln + qd] $F(4, 7) = 1.86$
T × G	T × G	T × G × P	T × G × H	T × G × P × H	T × G × P × H[ln + qd]
T[-1 0 1]	$F(2, 10) = 2.86$ T[-1 0 1]	$F(4, 20) = 0.92$ T[-1 0 1] × P	$F(8, 16) = 0.80$ T[-1 0 1] × H	$F(16, 8) = 0.33$ T[-1 0 1] × P × H	$F(8, 16) = 0.75$ T[-1 0 1] × P × H[ln + qd]
	$F(1, 10) = 6.05$	$F(2, 9) = 9.66$	$F(4, 7) = 0.38$	$F(8, 3) = 0.19$	$F(4, 7) = 0.83$

Note. The multivariate tests are based on Pillai's statistic. P = Phase; H = Hour; T = Treatment; G = Gender.

The Phase × Hour effect has 8 degrees of freedom. Thus, eight contrast variables must be defined. Now direct products become even more helpful:

$$\begin{aligned}
 P1 \times H1 &= [-1 \ 1 \ 0] \otimes [-2 \ -1 \ 0 \ 1 \ 2]y \\
 P1 \times H2 &= [-1 \ 1 \ 0] \otimes [2 \ -1 \ -2 \ -1 \ 2]y \\
 &\vdots \\
 P2 \times H4 &= [-1 \ 0 \ 1] \otimes [1 \ -4 \ 6 \ -4 \ 1]y.
 \end{aligned}$$

Finally, the AVERAGE variable is AVERAGE = [1/3 1/3 1/3] ⊗ [1/5 1/5 1/5 1/5 1/5]y.

Table 8 summarizes various tests associated with this design. (Once again, we included *F* statistics so that you can practice by replicating some of these analyses.) The results contained within the first four rows and first four columns should be self-explanatory. The last row gives results pertaining to several subeffects involving the comparison of the Control treatment and Treatment B.

The last column of results requires further elaboration. It contains results related to various subeffects involving the Phase × Hour interaction effect and related higher order interactions. In many applications of trend analysis, the lower order trend contrasts are believed to account for most of the meaningful variation in the data (the signal), whereas the higher order trend contrasts account for variation that is mostly randomness (the noise). For this example, we limited the Hour effect to just the linear and quadratic trends, pre-

tending that the population means are free of cubic and quartic trends. Using just  $P1 \times H1$ ,  $P1 \times H2$ ,  $P2 \times H1$ , and  $P2 \times H2$  simplifies the analysis and increases the statistical power, as long as the cubic and quartic variables really are dominated by noise. This strategy should be used regularly in repeated measures analyses involving many quantitative levels. Even if the levels are not quantitative, the researcher should define the smallest set of contrast variables that will reflect the hypothesized effects. This logic is no different from that of using parsimonious multiple-regression models.

### Designs With Several Within-Subjects Factors

When designs with more than two within-subject factors are encountered, direct products can greatly simplify the task of constructing the appropriate contrast variables. Consider a design with three within-subjects factors: *A* with 2 levels, *B* with 3 levels, and *C* with 5 levels. Each subject would then have 30 observations,  $y_{ijk}$ , which can be organized into the 30-element vector:  $y' = [y_{111} \ y_{112} \ \dots \ y_{115} \ y_{121} \ \dots \ y_{235}]$ . Contrast variables for this design can be defined by  $d = [a_1 \ a_2] \otimes [b_1 \ b_2 \ b_3] \otimes [c_1 \ c_2 \ c_3 \ c_4 \ c_5]y$ , where  $[a_1 \ a_2]$  is the contrast operating on the *A* factor and so on. For example, the *AB* interaction contrast variables can be defined as

$$AB1 = [1 \ -1] \otimes [1 \ -1 \ 0] \otimes [1 \ 1 \ 1 \ 1]y$$

$$AB2 = [1 \ -1] \otimes [1 \ 0 \ -1] \otimes [1 \ 1 \ 1 \ 1]y.$$

All the other contrast variables can also be constructed using this scheme. Then, just like before, one uses such sets of contrast variables as the dependent measures in a MANOVA that handles the between-subjects parts of the design.

### Familywise Testing With Bonferroni and Scheffé-Type Methods

Often it is desirable to place limits on the chance of one or more Type I errors occurring within a specific family of related hypothesis tests. There are many ways to control such familywise error rates in MANOVA. For the most part, we concur with Bird and Hadzi-Pavlovic (1983): Two acceptable, general, and straightforward ways are the Bonferroni and Scheffé procedures. Both of these methods may be used within MANOVA-based repeated measures analyses. But, as explained later, these two methods have different definitions of familywise.

#### Bonferroni Method

The Bonferroni is a general method that is applicable to any set of preplanned tests: It is not a method to test hypotheses that are suggested post hoc by the data.

Performing a Bonferroni test is straightforward. Suppose one's family of tests is composed of  $L$  individual preplanned tests. The familywise error rate,  $\alpha_{FW}$ , will be maintained if each of the  $L$  individual tests is conducted at the  $\alpha_{FW}/L$  level. For example, if one wanted to limit the familywise error rate to  $\alpha_{FW} = .05$  over three individual tests, those individual tests should be conducted at the  $.05/3 = .0167$  level of significance.

A slight improvement to the Bonferroni test was suggested by Šidák (1967). Instead of using  $\alpha_{FW}/L$  for each test, one uses  $1 - (1 - \alpha_{FW})^{1/L}$ . For  $L = 3$  and  $\alpha_{FW} = .05$ , each test would be conducted at the .0170 level, only trivially different from the Bonferroni level of .0167. However, if a Bonferroni test just misses being significant, there is nothing wrong with using the more accurate Šidák level, which is always slightly greater.

Carrying out the Bonferroni method is now quite easy, because most statistical software reports significance levels ( $p$  values) to several digits. The Bonferroni method is especially useful when one wishes to control the Type I error rate over just a few (say  $L \leq 5$ ) preplanned tests.

#### Scheffé-Type Method

The Scheffé-type method uses Roy's (1957, chap. 14) general method for constructing simultaneous confidence intervals in the multivariate linear model. The familiar Sheffé test for univariate ANOVA is a special case. The notion of family for this method is more abstract, so for our purposes it will be best to begin by working with a concrete example. Consider again the Group $[-2 \ 1 \ 1] \times$  Time $[-1 \ 1 \ 0]$  effect from the design with three groups (control, Treatment A, Treatment B) and three measurements (pretest, posttest, follow-up).

For Scheffé-type testing, a family of tests consists of all possible subeffects of a given parent effect. If the parent is a 1-degree-of-freedom effect, then it has no subeffects. If the parent is more than 1 degree of freedom, it has an uncountably infinite number of subeffects. The Scheffé-type procedure allows us to test as many of these subeffects as we desire, without increasing the chance that a single Type I error will occur within that protected family.

What is the parent effect for Group $[-2 \ 1 \ 1] \times$  Time $[-1 \ 1 \ 0]$ ? Usually we would say it is the Group  $\times$  Time interaction (this is discussed in detail later). Besides Group $[-2 \ 1 \ 1] \times$  Time $[-1 \ 1 \ 0]$ , there are countless other subeffects of Group  $\times$  Time. Among them are Group $[-1 \ 0 \ 1] \times$  Time, Group  $\times$  Time $[-1 \ 1 \ 0]$ , and Group $[-1 \ 0 \ 1] \times$  Time $[-1 \ 0 \ 1]$ . Even Group $[-9 \ 1 \ 8] \times$  Time $[-.16 \ .01 \ .15]$  is a subeffect. It is important to realize that because all subeffects are included in the family, we can look at the data to choose selectively the subeffects of the parent effect that are most likely to be significant. It can be shown that if the parent effect is significant according to Roy's statistic, then there must be significant subeffects according to the Scheffé-type method. (Bear in mind, however, that a significant subeffect may not

be a meaningful subeffect.) In contrast, if the parent effect is not significant according to Roy's statistic, none of the parent's subeffects are significant according to the Scheffé-type method. We know that Group  $\times$  Time is a fruitful parent effect, because, as noted earlier, its largest root statistic,  $\theta_1 = .640$ , exceeded its critical value,  $\theta [s = 2, m = -.5, n = 5, \alpha = .05] = .498$ .

Any subeffect with 1 degree of freedom can be Scheffé tested by first computing the preplanned, univariate  $F$  statistic as previously shown, and then comparing that  $F$  statistic to the Scheffé critical value,

$$F_S = (N - g)(\theta[s, m, n, \alpha_{FW}]) / (1 - \theta[s, m, n, \alpha_{FW}]),$$

where the values of  $s, m, n$  correspond to those of the parent effect. When  $s = 1$ , we use

$$F_S = (N - g)(v_1/v_2)F[v_1, v_2, \alpha_{FW}],$$

where  $v_1 = Bdf \cdot Wdf, v_2 = N - g - Wdf + 1$ , and  $F[v_1, v_2, \alpha_{FW}]$  is a critical value from the  $F[v_1, v_2]$  distribution.

All subeffects of the Group  $\times$  Time interaction have the same critical value. Here,  $N = 16, g = 3$ , and  $\theta[2, -.5, 5, .05] = .498$  lead to  $F_S = (16 - 3)(.498)/(1 - .498) = 12.90$ . Glancing at Table 6,  $F = 8.00$  for Group  $[-2 \ 1 \ 1] \times$  Time  $[-1 \ 1 \ 0]$ , which is not significant. Because  $F = 18.86$  for Group  $[-2 \ 1 \ 1] \times$  Time  $[-1 \ 0 \ 1]$ , it is significant.

Defining the parent effect must be a preplanned decision. In doing so, researchers should understand that using more specific parents leads to lower Scheffé critical values, which often yield more powerful Scheffé tests. For example, we could define Group  $[-2 \ 1 \ 1] \times$  Time as a parent effect, thereby producing a family of contrasts which is a subset of the Group  $\times$  Time family. With this parent,  $Bdf = 1$  and  $Wdf = 2$ , so that  $s = 1$ . After finding  $F[2, 12, .05] = 3.89$  in an ordinary  $F$  table, we get  $F_S = (16 - 3)(2/12)(3.89) = 8.42$ . This is appreciably less than the previous critical value, 12.90. If the power for a Group  $[-2 \ 1 \ 1] \times$  Time  $[t_1 \ t_2 \ t_3]$  contrast is .75 using  $F_S = 12.90$ , it is .91 using  $F_S = 8.42$  (noncentrality  $\lambda = 19.0$ ). Scheffé testing is often criticized for being unpowerful: Researchers and statisticians tend

to use parent effects that are too general. Through judicious preplanning of more specific parent effects, the power of Scheffé testing can be markedly increased. Of course, this strategy decreases the generality of the Type I error protection.

For a given parent effect, there is an optimal subeffect, defined by one between-subjects contrast and one within-subjects contrast variable, that maximizes the  $F$  statistic for a given data set. Using optimal subeffects as starting places can help researchers design meaningful candidates for significant Scheffé-type contrasts. This strategy, which is too involved and technical to develop here, was discussed by Bird and Hadzi-Pavlovic (1983), but they limited their discourse to balanced one-way MANOVA designs. In the Appendix, we modify their equations to handle applications to repeated measures designs with unequal group sizes.

How large can a family of Bonferroni contrasts be before  $\alpha_{FW}/L$  is smaller than the effective per-contrast error rate for the corresponding Scheffé-type method? Table 9 gives some calculations for a case with  $N - g = 107$  and  $\alpha_{FW} = .05$ . To illustrate, consider a parent with  $Bdf = 2$  and  $Wdf = 3$ : The Scheffé critical value sets a per-contrast error rate that is approximately .05/42. Table 9 dramatizes the fact that one can test many preplanned contrasts with Bonferroni protection before switching to the method that protects all subeffect contrasts. It also shows how weak the Scheffé-type method becomes as  $Bdf$  and  $Wdf$  increase.

In closing, we should state that many analyses can profit by the careful use of both types of familywise strategies. As Bird and Hadzi-Pavlovic recommended, some families can even be broken up into  $L$  subfamilies, which can then be tested with the Scheffé-type procedure operating at the  $\alpha_{FW}/L$  subfamilywise error rate.

#### Assumptions of the MANOVA Approach

For completeness, let us now state the mathematical assumptions underlying the MANOVA approach and briefly summarize how the method is affected by violations of those assumptions. Consider a test involving  $g$  groups and  $Wdf$  contrast variables. When

Table 9  
Comparing Bonferroni and Scheffé-Type Tests on  
the Basis of Break-Even  $L$  Values

Bdf for parent effect	Wdf for parent effect			
	1	2	3	6
1	1	4	11	174
2	4	14	48	1060
3	10	42	162	4847
6	125	716	3474	136696

Note. Wdf = degrees of freedom for within-subjects part of test; Bdf = degrees of freedom for between-subjects part of test. Values computed for  $(N - g) = 107$  and  $\alpha_{FW} = .05$ .

using the MANOVA method, let us assume that those variables have a multivariate normal distribution. The Wdf variances of the contrast variables and their Wdf(Wdf - 1)/2 intercorrelations may be of any structure, but we assume that this structure is homogeneous across the groups. Finally, we assume that each subject's observations on the contrast variables are independent of the other subjects' observations. (The traditional mixed-model repeated measures analysis carries these same assumptions, but it also puts the sphericity restrictions on the structure of the variances and correlations.)

Statisticians have investigated how MANOVA tests behave when their assumptions are violated. (See Olsen, 1974, 1976, and the references therein.) In general, MANOVA's robustness parallels ANOVA's robustness. Multivariate normality can be violated to a significant degree without seriously affecting the validity of the  $p$  values or the powers of the tests. Not having homogeneity of variances and correlations is problematic if the sample sizes of the independent groups are unequal, but for equal or nearly equal group sizes, MANOVA is acceptably robust to this assumption. If intersubject independence is violated, serious consequences usually result.

We already mentioned that the Pillai statistic may be more robust than its closest competitors, Wilks's lambda and Hotelling's statistics. There is also valid evidence that Roy's largest root test is much less robust to violations of normality and homogeneity of variances and correlations. This fact generates some uneasiness about the Scheffé method-

ology, but there is no alternative to Sheffé testing that has its generality, statistical power, and ease of use. For much more on this matter, see Bird and Hadzi-Pavlovic (1983).

We do not recommend using hypothesis tests to pretest for the various assumptions. Hypothesis tests are designed to detect violations from a perfect null situation; they generally do not indicate when a particular assumption is violated in such a way that it becomes problematic for a particular procedure. When sample sizes are small, such pretesting lacks the power to pick up problematic violations of assumptions. When sample sizes are large, such pretesting has enough power to pick up inconsequential violations of assumptions. Often the pretest is less robust than the main procedure. This is the case for Box's  $M$  test for the homogeneity of the variances and correlations, which is not robust to nonnormality (Olson, 1974, p. 906).

## Conclusion

In summary, the MANOVA approach for repeated measures affords a robust, flexible alternative to the traditional mixed-model analysis. Further, given the properties of the data generated by most repeated measures studies, the researcher sacrifices little power by using the MANOVA approach. In fact, because the researcher is easily able to carry out specific tests of the experimental hypotheses, the MANOVA approach is actually more powerful than the traditional approach in many cases.

As we hope we demonstrated through the examples in this article, the MANOVA approach is not particularly complex, nor does it require a high degree of statistical sophistication on the part of the researcher, given the ready availability of suitable statistical software in most research settings. Further, once the fundamental technique of the MANOVA approach is mastered, the analysis can be expanded and applied to any fixed-effects repeated measures design, irrespective of complexity. Overall, the power, versatility, and precision of the MANOVA approach is unsurpassed by any competing method. We encourage you to explore the application of



the MANOVA approach in your own research and hope that this article makes that exploration somewhat less precarious.

### References

- Bird, K. D., & Hadzi-Pavlovic, D. (1983). Simultaneous test procedures and the choice of a test statistic in MANOVA. *Psychological Bulletin*, *93*, 167-178.
- Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.
- Boik, R. J. (1981). A priori tests in repeated measures designs: Effects of nonsphericity. *Psychometrika*, *46*, 241-255.
- Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: II. Effect of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics*, *25*, 484-498.
- Burns, P. R. (1984). *SPSS-600 MANOVA Version IV update manual*. Evanston, IL: Northwestern University, Vogelback Computing Center.
- Collier, R. O., Baker, F. B., Mandeville, G. K., & Hayes, T. F. (1967). Estimates of test size for several procedures based on conventional ratios in the repeated measure design. *Psychometrika*, *32*, 339-353.
- Davidson, M. L. (1972). Univariate versus multivariate tests in repeated measures experiments. *Psychological Bulletin*, *77*, 446-452.
- Harris, R. J. (1975). *A primer of multivariate statistics*. New York: Academic Press.
- Hearne, E. M., III, Clark, G. M., & Hatch, J. P. (1983). A test for serial correlation in univariate repeated-measures analysis. *Biometrics*, *39*, 237-243.
- Herr, D. G., & Gaebelin, J. (1978). Nonorthogonal two-way analysis of variance. *Psychological Bulletin*, *85*, 207-216.
- Huynh, H. (1978). Some approximate tests in repeated measures designs. *Psychometrika*, *43*, 1582-1589.
- Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in the randomized block and split-plot designs. *Journal of Educational Statistics*, *1*, 69-82.
- Huynh, H., & Feldt, L. S. (1980). Performance of traditional *F* tests in repeated measures designs under variance heterogeneity. *Communications in Statistics: Series A*, *9*, 61-74.
- Huynh, H., & Mandeville, G. K. (1979). Validity conditions in repeated measures designs. *Psychological Bulletin*, *86*, 964-973.
- Keppel, G. (1982). *Design and analysis: A researcher's handbook* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Keselman, H. J., Rogan, J. C., Mendoza, J. L., & Breen, L. L. (1980). Testing the validity conditions of repeated measures *F* tests. *Psychological Bulletin*, *87*, 479-481.
- Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Wadsworth.
- Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. *Journal of Educational Statistics*, *5*, 269-287.
- McCall, R. B., & Appelbaum, M. I. (1973). Bias in the analysis of repeated-measures designs: Some alternative approaches. *Child Development*, *44*, 401-415.
- Mendoza, J. L., Toothaker, L. E., & Nicewander, W. A. (1974). A Monte Carlo comparison of the univariate and multivariate methods for the group by trials repeated measures design. *Multivariate Behavioral Research*, *9*, 165-177.
- Mitzel, H. C., & Games, P. A. (1981). Circularity and multiple comparisons in repeated measures designs. *British Journal of Mathematical and Statistical Psychology*, *34*, 253-259.
- Myers, J. L. (1979). *Fundamentals of experimental design* (3rd ed.). Boston: Allyn & Bacon.
- Nanda, D. N. (1950). Distribution of the sum of roots of a determinantal equation under a certain condition. *Annals of Mathematical Statistics*, *21*, 432-439.
- Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, *69*, 894-908.
- Olson, C. L. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, *83*, 579-586.
- Rogan, J. C., Keselman, H. J., & Mendoza, J. L. (1979). Analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology*, *32*, 269-286.
- Roy, S. N. (1957). *Some aspects of multivariate analysis*. New York: Wiley.
- Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*, 626-633.
- Srivastava, M. S., & Khatri, C. G. (1979). *An introduction to multivariate statistics*. New York: Elsevier-North Holland.
- Timm, N. H. (1980a). The analysis of nonorthogonal MANOVA designs employing a restricted full rank multivariate linear model. In R. P. Gupta (Ed.), *Multivariate statistical analysis* (pp. 257-273). New York: Elsevier-North Holland.
- Timm, N. H. (1980b). Multivariate analysis of variance of repeated measurements. In P. R. Krishnaiah (Ed.), *Handbook of statistics. Volume I: Analysis of variance* (pp. 41-87). New York: Elsevier-North Holland.
- Wallenstein, S., & Fleiss, J. L. (1979). Repeated measures analysis of variance when the correlations have a certain pattern. *Psychometrika*, *44*, 229-233.
- Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

## Appendix

## Optimal Contrasts for Unbalanced Repeated Measures Designs

Similar to the notational style of Bird and Hadzi-Pavlovic (1983), let  $\mu$  be the  $g \times p$  matrix of population means, where  $g$  is the total number of independent groups and  $p$  is the total number of repeated measures. Let  $M$  be the  $g \times p$  matrix of sample means and  $D$  be the  $g \times g$  diagonal matrix of the sample sizes:  $\text{diag}(D) = (n_1 \ n_2 \ \dots \ n_g)$ . If  $C$  is the  $Bdf \times g$  matrix defining the between-subjects effect and  $W$  is the  $p \times Wdf$  matrix defining the within-subjects effect, the hypothesis for the parent effect is  $H_0: C\mu W = 0$ . When one uses  $W$  to compute the  $Wdf$  contrast variables, which then are used in a MANOVA test of the  $C$  effect, one obtains Roy's largest root,  $\theta$ , [or  $\lambda_1 = \theta_1/(1 - \theta_1)$ ], and the first set of raw discriminant weights,  $\mathbf{a}_1 = (a_1 \ a_2 \ \dots \ a_{Wdf})$ .

The optimal single-degree-of-freedom subeffect of  $H_0: C\mu W = 0$  is  $H_0: \mathbf{c}'\mu\mathbf{w} = 0$ , where  $\mathbf{w} = W\mathbf{a}_1$  and  $\mathbf{c} = C'(CD^{-1}C')^{-1}CM\mathbf{w}$ . SPSSX-MANOVA gives  $\mathbf{a}_1$ , and the updated version (Burns, 1984) will now give  $\mathbf{c}$ .

The ordinary  $F$  statistic obtained by testing  $H_0: \mathbf{c}'\mu\mathbf{w} = 0$  is equal to  $(N - g)\theta_1/(1 - \theta_1)$ , making it easy to see why this optimized  $F$  is tested using a critical value based on Roy's largest root. Because this critical value protects this optimal contrast, it simultaneously protects all other subeffects of  $H_0: C\mu W = 0$ .

Received January 31, 1984

Revision received June 29, 1984 ■