Psychological Bulletin
1995, Vol. 117, No. 3, 547-560

0033-2909/95/$3.00

Approximate Degrees of Freedom Tests: A Unified Perspective on Testing for Mean Equality

Lisa M. Lix and H. J. Keselman
University of Manitoba

This article presents a statistic for tests of mean equality in between-subjects and within-subjects designs when variances are heterogeneous. The approximate degrees of freedom statistic of S. Johansen (1980) can be used to test main and interaction effects, as well as multiple comparison hypotheses related to these effects. Thus, researchers need only be familiar with a single statistic, rather than the many statistics that have been defined in the literature, to perform these tests of significance. Also included is a computer program to obtain a numerical solution.

If valid tests of mean equality are to be obtained with classical procedures (e.g., Student's t statistic, Snedecor's F statistic), homogeneity of variance assumptions must be satisfied. Exactly which variances are assumed to be equal depends on the type of research design. For example, when mean equality in univariate between-subjects designs is being tested, this assumption implies the equality of the group (or cell) population variances, whereas in the multivariate cases of these designs, the assumption refers to equality of the population variance–covariance matrices. In within-subjects designs containing at least one between-subjects grouping variable, valid tests of the equality of repeated measures means rest in part on the assumption that the between-subjects covariance matrices of a set of orthonormalized repeated measures contrast variables are equal.

It is well known that the classical tests of mean equality can become seriously biased when homogeneity of variance assumptions are not satisfied, particularly when the design is unbalanced (i.e., unequal group [cell] sizes; Maxwell & Delaney, 1990, pp. 723–724). Specifically, the classical procedures can become considerably conservative or liberal, with rates of Type I error lower than .01 or higher than .70, respectively, for a .05 significance level (e.g., Keselman & Keselman, 1988; Milligan, Wong, & Thompson, 1987).

There is a long history in the statistical literature, going back to Behrens (1929) and Fisher (1935), concerning the problem of testing for mean equality when equality of variances cannot be assumed (Roth, 1988). One solution to the problem that has been frequently suggested in this literature is to use a nonpooled estimate of error variance in computing a test statistic and to incorporate the variances and group sizes into an approximate estimate for error degrees of freedom. Welch (1938, 1947, 1951) provided such an approximate degrees of freedom (ADF) solution for tests of mean equality, first in the two-group between-subjects design and then in the multigroup extension of this design. His procedure has also been applied and examined within the context of factorial between-subjects designs. Specifically, Keselman, Carriere, and Lix (in press) used an ADF statistic for examining omnibus main and interaction effects in unbalanced factorial designs, whereas Hsiung and Olejnik (1994a) defined such a test for pairwise comparisons of the marginal main effect means. In addition, Algina, Oshima, and Tang (1991) defined an ADF test for the one-way between-subjects design in which there are multiple dependent variables (see also Tang & Algina, 1993). For designs involving repeated measurements, Keselman, Carriere, and Lix (1993); Keselman, Keselman, and Shaffer (1991); and Lix and Keselman (in press) defined and examined Welch-type statistics to test omnibus main and interaction effect hypotheses, multiple marginal pairwise comparison hypotheses, and multiple interaction contrast hypotheses, respectively.

Empirical results indicate that, under many conditions likely to be encountered by behavioral scientists, these nonpooled statistics with their approximate estimates of error degrees of freedom provide relatively robust tests for mean equality when homogeneity requirements are not satisfied (e.g., Algina et al., 1991; Keselman et al., 1991, 1993, in press; Lix & Keselman, in press; Tang & Algina, 1993). Moreover, the power of an ADF test is generally comparable to that of its traditional counterpart, even when the homogeneity assumption is satisfied (Brown & Forsythe, 1974; Dijkstra & Werter, 1981; Hsiung & Olejnik, 1994b; Keselman et al., in press; Roth, 1983; Tomarken & Serlin, 1986; Wilcox, Charlin, & Thompson, 1986). Furthermore, ADF tests may be more powerful than the traditional tests when both variances and group (cell) sizes are unequal and positively paired, so that the largest variance is associated with the group (cell) having the largest number of observations (Keselman et al., in press).

However, it should be noted at the outset that ADF tests are not always robust. The empirical literature generally indicates that in univariate designs, when unequal variances occur in combination with unequal group (cell) sizes so that the largest variance is paired with the smallest group (cell) size (i.e., a negative pairing) and the data are also extremely skewed in shape

This research was supported by a Social Sciences and Humanities Research Council (SSHRC) of Canada doctoral fellowship (Award 752-92-1628) and an SSHRC research grant (Award 410-92-0430).

Correspondence concerning this article should be addressed to Lisa M. Lix or to H. J. Keselman, Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2. Electronic mail may be sent via Internet to kesel2@ccm.umanitoba.ca (Lisa M. Lix) or to kesel@ccm.umanitoba.ca (H. J. Keselman).

(Micceri, 1989), ADF tests cannot limit the probability of a Type I error to the nominal level of significance (Keselman et al., in press). Moreover, in multivariate designs, if the ratio of the smallest group (cell) size to the number of dependent variables or repeated measurements is small, covariance matrices are unequal, and the data are skewed, the Type I error rates may become severely inflated (Algina et al., 1991; Keselman et al., 1993). Nonetheless, ADF tests do, in many situations likely to be encountered by behavioral scientists, provide reasonably good Type I error control and a powerful alternative to traditional tests of mean equality. In addition, both parametric and nonparametric alternatives to ADF procedures suffer from their own weaknesses and either do not test the same hypotheses (Sawilowsky, 1990) or do not control the Type I error rate under the combined effects of nonnormality and heteroscedasticity. For example, James's (1954) second-order solution has been found to provide better Type I error control than the ADF test (Wilcox, 1988), but it also becomes liberal when variances and group sizes are negatively paired and the data are nonnormal (Algina et al., 1991; Hsiung & Olejnik, 1994b; Lix & Keselman, 1994).

Although ADF nonpooled statistics provide an attractive alternative to the classical procedures for testing equality of means, the apparent diversity of statistical tests appearing in the literature and of sources in which these procedures appear might dissuade potential users from adopting this approach. However, all of the previously enumerated ADF test statistics are obtainable from the general formulation of Johansen (1980). Accordingly, the purpose of the present article is to present this unified perspective and to illustrate its application in unbalanced between-subjects and within-subjects designs containing one or more dependent variables.

Furthermore, we present a computer program for obtaining numerical results using this general solution for tests of mean equality in designs that do not contain quantitative covariates. This program requires only that the user enter the data, the number of observations per group (cell), and the coefficients of one or more contrast matrices that represent the hypothesis of interest. For completeness, and to show how these three specifications relate to the ADF solution, we first present, in abbreviated form, the mathematical underpinnings of the approach. However, we emphasize that a complete understanding of this presentation is not necessary for applied researchers to use the accompanying computer program.

A General ADF Test Statistic

Matrix notation is used here to develop a general approach for testing hypotheses of mean equality using an ADF solution. The multivariate perspective is considered first; the univariate model is a special case of the multivariate.

Consider the general linear model (GLM; see Timm, 1975, for a more detailed discussion of the GLM approach)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}, \tag{1}$$

where $\mathbf{Y}$ is an $N \times p$ matrix of scores on p dependent variables or p repeated measurements, N is the total sample size, $\mathbf{X}$ is an $N \times r$ design matrix consisting entirely of zeros and ones with $\text{rank}(\mathbf{X}) = r$, $\boldsymbol{\beta}$ is an $r \times p$ matrix of nonrandom parameters (i.e., population means), and $\boldsymbol{\xi}$ is an $N \times p$ matrix of random error components.

Let $\mathbf{Y}_j$ ($j = 1, \ldots, r$) denote the submatrix of $\mathbf{Y}$ containing the scores associated with the n_j subjects in the jth group (cell). It is assumed that the rows of $\mathbf{Y}_j$ are independently and normally distributed, with mean vector $\boldsymbol{\beta}_j$ and variance–covariance matrix $\boldsymbol{\Sigma}_j$ [i.e., i.d. $N(\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_j)$], where $\boldsymbol{\beta}_j = (\mu_{j1} \cdots \mu_{jp})$, the jth row of $\boldsymbol{\beta}$, and $\boldsymbol{\Sigma}_j \neq \boldsymbol{\Sigma}_{j'}$ ($j \neq j'$). Specific formulas for estimating $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_j$, as well as an elaboration of $\mathbf{Y}_j$, are provided in Appendix A.

The general linear hypothesis is

$$H_0: \mathbf{R}\boldsymbol{\mu} = \mathbf{0}, \tag{2}$$

where $\mathbf{R} = \mathbf{C} \otimes \mathbf{U}^{\mathrm{T}}$; $\mathbf{C}$ is a $df_{\mathrm{C}} \times r$ matrix that controls contrasts on the between-subjects effect(s), with $\text{rank}(\mathbf{C}) = df_{\mathrm{C}} \leq r$; $\mathbf{U}$ is a $p \times df_{\mathrm{U}}$ matrix that controls contrasts on the within-subjects effect(s), with $\text{rank}(\mathbf{U}) = df_{\mathrm{U}} \leq p$; $\otimes$ is the Kronecker or direct product function; and superscript T is the transpose operator. For multivariate between-subjects designs, $\mathbf{U}$ is an identity matrix of dimension p (i.e., $\mathbf{I}_p$). The $\mathbf{R}$ contrast matrix has $df_{\mathrm{C}} \times df_{\mathrm{U}}$ rows and $r \times p$ columns. In Equation 2, $\boldsymbol{\mu} = \text{vec}(\boldsymbol{\beta}^{\mathrm{T}}) = (\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_r)^{\mathrm{T}}$. In other words, $\boldsymbol{\mu}$ is the column vector with $r \times p$ elements obtained by stacking the columns of $\boldsymbol{\beta}^{\mathrm{T}}$. The $\mathbf{0}$ column vector is of order $df_{\mathrm{C}} \times df_{\mathrm{U}}$.

To illustrate, in a design containing a between-subjects factor with four levels and a within-subjects factor with three levels, $\mathbf{C}$ and $\mathbf{U}$ matrices of possible interest are

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix} \text{ and } \mathbf{U} = \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Here, the rows of $\mathbf{C}$ represent a set of three linearly independent contrasts among the levels of the between-subjects factor, whereas the columns of $\mathbf{U}$ form a set of two linearly independent contrasts among the levels of the within-subjects factor (Kirk, 1982, pp. 226, 784–785). The Kronecker product, $\mathbf{C} \otimes \mathbf{U}^{\mathrm{T}}$, is

$$\mathbf{R} = \begin{bmatrix} (1)\mathbf{U}^{\mathrm{T}} & (-1)\mathbf{U}^{\mathrm{T}} & (0)\mathbf{U}^{\mathrm{T}} & (0)\mathbf{U}^{\mathrm{T}} \\ (1)\mathbf{U}^{\mathrm{T}} & (0)\mathbf{U}^{\mathrm{T}} & (-1)\mathbf{U}^{\mathrm{T}} & (0)\mathbf{U}^{\mathrm{T}} \\ (1)\mathbf{U}^{\mathrm{T}} & (0)\mathbf{U}^{\mathrm{T}} & (0)\mathbf{U}^{\mathrm{T}} & (-1)\mathbf{U}^{\mathrm{T}} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

and has six linearly independent rows.

The generalized test statistic provided by Johansen (1980) is

$$T_{\mathrm{WJ}} = (\mathbf{R}\hat{\boldsymbol{\mu}})^{\mathrm{T}}(\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^{\mathrm{T}})^{-1}(\mathbf{R}\hat{\boldsymbol{\mu}}), \tag{3}$$

where $\hat{\boldsymbol{\mu}}$ estimates $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\Sigma}} = \text{diag}(\hat{\boldsymbol{\Sigma}}_1/n_1 \cdots \hat{\boldsymbol{\Sigma}}_r/n_r)$, a block matrix with diagonal elements $\hat{\boldsymbol{\Sigma}}_j/n_j$. This statistic, divided by a constant, c, approximately follows an F distribution with $\nu_1 = df_{\mathrm{C}} \times df_{\mathrm{U}}$ and $\nu_2 = \nu_1(\nu_1 + 2)/(3A)$, where $c = \nu_1 + 2A - (6A)/(\nu_1 + 2)$. The formula for the statistic A is provided in Appendix A.

When $p = 1$, that is, for a univariate model, the elements of $\mathbf{Y}_j$ are assumed to be independently and normally distributed with mean μ_j and variance σ_j^2 [i.e., i.d. $N(\mu_j, \sigma_j^2)$]. To test the general linear hypothesis, $\mathbf{C}$ has the same form and function as for the multivariate case, but $\mathbf{U} = 1$. With respect to the computation of T_{WJ}, $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1 \cdots \hat{\mu}_r)^{\mathrm{T}}$ and $\hat{\boldsymbol{\Sigma}} = \mathrm{diag}(\hat{\sigma}_1^2/n_1 \cdots \hat{\sigma}_r^2/n_r)$. (See Appendix A for further details of the univariate model.)

Obtaining Numerical Results Using an ADF Solution

One constraint to applied researchers adopting ADF nonpooled statistics is the lack of availability of these statistics in commercial statistical packages. For univariate designs, the BMDP (Dixon, Brown, Engelman, Hill, & Jennrich, 1988), SAS (SAS Institute, 1989b), and SPSS (Norusis, 1990) programs all produce results for the two-group between-subjects design (i.e., the nonpooled t test). However, only BMDP allows the user to compute such a statistic for the one-way between-subjects design. Intermediate results (i.e., group means and variances, or covariance matrices in the case of multivariate or repeated measures designs) can be obtained from all three packages, but researchers may find the final calculations tedious and time consuming, particularly if they must be done by hand.

In response to the lack of availability of programs containing ADF tests, Chenier and Seaman (1987) devised a PASCAL program to compute a nonpooled solution for univariate between-subjects designs containing up to three factors. Subsequently, other programs were developed for either ADF statistics (Keselman et al., 1993, in press) or other alternatives to classical procedures, such as James's (1954) second-order procedure (Hsiung, Olejnik, & Oshima, 1994; Oshima & Algina, 1992). However, the lack of generalizability of existing programs impedes their usefulness. As well, because these programs compute tests only of omnibus effects, researchers must rely on hand computations to produce results for single or multiple degrees of freedom contrasts, often the most informative component of a statistical analysis.

Appendix B contains a module of statements written in the SAS/IML (SAS Institute, 1989a) programming language that can be used to obtain numerical results for the general ADF solution described in the previous section. Tests of omnibus main effects or interaction effects, or both, may be performed, in addition to tests of contrasts, making the program versatile for a variety of research designs and hypotheses of potential interest to applied researchers.

The program requires as input $\mathbf{Y}$, $\mathbf{C}$, and $\mathbf{NX}$, the latter being a $1 \times r$ vector containing the number of observations in each group or cell (i.e., the n_js). It is assumed that the orders of entry for $\mathbf{NX}$ and $\mathbf{Y}$ correspond so that the first n_1 rows of $\mathbf{Y}$ correspond to the first element of $\mathbf{NX}$, the next n_2 elements of $\mathbf{Y}$ correspond to the second element of $\mathbf{NX}$, and so on. By default, $\mathbf{U} = \mathbf{I}_p$, but for within-subjects designs the user must specify the elements of $\mathbf{U}$. The module computes $\mathbf{R}$ from the designated $\mathbf{C}$ and $\mathbf{U}$ matrices. The module is invoked with a RUN WJGLM statement and supplies as output $\mathbf{R}$, $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$, T_{WJ}/c, ν_1, ν_2, and the associated p value.

The following section is devoted to illustrating how to test hypotheses of mean equality on between-subjects effects, within-subjects effects, or both in various research designs from the GLM perspective using the statistic T_{WJ} of Equation 3 and the SAS/IML (SAS Institute, 1989a) program developed for its application. In many of these examples, pairwise contrasts are computed to identify specific treatment effects. The tests are restricted to this set for illustration purposes and correspond to applications of ADF solutions presented in the literature.

Applications of the General ADF Solution

One-Way Between-Subjects Designs

In a between-subjects experiment with n_j subjects ($\Sigma_{j=1}^{J} n_j = N$) in each of J groups, let Y_{ij} denote the score associated with the ith subject in the jth group ($j = 1, \ldots, J$; $i = 1, \ldots, n_j$). Adopting a cell means model (Maxwell & Delaney, 1990, pp. 86–95),

$$Y_{ij} = \mu_j + \epsilon_{ij}, \tag{4}$$

where μ_j is the jth population mean and ϵ_{ij} is the random error term associated with the ijth observation. The Y_{ij}s are assumed to be i.d. $N(\mu_j, \sigma_j^2)$ variates, with $\hat{\mu}_j$ and $\hat{\sigma}_j^2$, respectively, representing the jth sample mean and unbiased variance.

When the classical assumptions of independence, normality, and variance homogeneity are met, the usual analysis of variance (ANOVA) F test provides the uniformly most powerful invariant and exact test of H_J: $\mu_1 = \cdots = \mu_J$. However, when variance homogeneity is not a tenable assumption, the F test is known to produce invalid results (Harwell, Rubinstein, Hayes, & Olds, 1992). Consequently, researchers may wish to routinely adopt a parametric alternative to the F test that does not rest on this assumption. Welch (1951) demonstrated such a solution for the one-way design.

The Welch (1951) statistic is

$$F(\mathrm{W}) = \frac{\sum_{j=1}^{J} w_j(\hat{\mu}_j - \hat{\mu}^*)^2/(J-1)}{1 + \dfrac{2(J-2)}{(J^2-1)} \sum_{j=1}^{J} \dfrac{(1 - w_j/W)^2}{n_j - 1}}, \tag{5}$$

where $w_j = n_j/\hat{\sigma}_j^2$, $\hat{\mu}^* = \Sigma_{j=1}^{J} w_j \hat{\mu}_j / W$, and $W = \Sigma_{j=1}^{J} w_j$. The test statistic is approximately distributed as an F variate and is referred to the critical value $F[(1-\alpha); (J-1), \nu_{F(\mathrm{W})}]$, the $(1-\alpha)$ centile of the F distribution, where

$$\nu_{F(\mathrm{W})} = \frac{J^2 - 1}{3 \sum_{j=1}^{J} \dfrac{(1 - w_j/W)^2}{n_j - 1}}. \tag{6}$$

In applying Welch's (1951) solution within the context of Johansen's (1980) general ADF approach and adopting the notation of Equation 1, $\mathbf{Y} = (Y_{ij})$, the $N \times 1$ vector with elements Y_{ij}, $\mathbf{X}$ is the $N \times J$ matrix with ones denoting presence in a group and zeros denoting absence, $\boldsymbol{\beta}^{\mathrm{T}} = (\mu_1 \cdots \mu_J)$, and $\boldsymbol{\xi} = (\epsilon_{ij})$. To test the general linear hypothesis of Equation 2, $\mathbf{R} = \mathbf{C} = \mathbf{C}_j$ because $\mathbf{U} = 1$. Thus, $\mathbf{R}$ is a $(J-1) \times J$ matrix for which the rows represent a set of linearly independent contrasts among the levels of the between-subjects factor. With respect to Equation 3, $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1 \cdots \hat{\mu}_J)^{\mathrm{T}}$ and $\hat{\boldsymbol{\Sigma}} = \mathrm{diag}(\hat{\sigma}_1^2/n_1 \cdots \hat{\sigma}_J^2/n_J)$.

Contrasts on the group means may be tested, with pairwise

contrasts frequently being of greatest interest to researchers. However, multiple comparison procedures that use a test statistic based on the usual pooled estimate of error variance are sensitive to variance heterogeneity (Dunnett, 1980; Games & Howell, 1976). In such situations, multiple comparison procedures that rely on a nonpooled test statistic, as described by Welch (1938) and Games and Howell (1976), are recommended (see also Maxwell & Delaney, 1990, pp. 146–150). To test $H_{jj'}$: $\mu_j = \mu_{j'}$, where $j \neq j'$, the nonpooled statistic is

$$t(\mathrm{W}) = \frac{\hat{\mu}_j - \hat{\mu}_{j'}}{\sqrt{\frac{\hat{\sigma}_j^2}{n_j} + \frac{\hat{\sigma}_{j'}^2}{n_{j'}}}}, \tag{7}$$

which is approximately distributed as a t variate with the critical value $t[(1 - \alpha/2); \nu_{t(\mathrm{w})}]$, the $(1 - \alpha/2)$ centile of Student's t distribution with degrees of freedom

$$\nu_{t(\mathrm{W})} = \frac{\left(\frac{\hat{\sigma}_j^2}{n_j} + \frac{\hat{\sigma}_{j'}^2}{n_{j'}}\right)^2}{\frac{(\hat{\sigma}_j^2/n_j)^2}{n_j - 1} + \frac{(\hat{\sigma}_{j'}^2/n_{j'})^2}{n_{j'} - 1}}. \tag{8}$$

The general ADF solution may also be used to conduct pairwise contrasts, eliminating the need to rely on the specific formulas detailed in Equations 7 and 8. In terms of the hypothesis of Equation 2, $\mathbf{R} = \mathbf{C} = \mathbf{c}_{jj'} = (c_1 \cdots c_J)$, the $1 \times J$ vector of coefficients that contrasts the jth and j'th means ($\Sigma_{j=1}^{J} c_j = 0$).

An example data set was taken from Glass and Hopkins (1984, p. 330), for a study of the effect of an anxiety reduction program on student test scores, to demonstrate the SAS/IML (SAS Institute, 1989a) program for the general solution in a one-way between-subjects design. Ten participants were randomly assigned to each of three anxiety reducing treatments: low, moderate, and high. However, to show how the program is applied to an unbalanced design, we have deleted the last four scores associated with the first condition.

Although researchers may use standard SAS/IML (SAS Institute, 1989a) notation to input the elements of the **C** matrix for testing the hypothesis of no overall group effect, formation of this matrix can be greatly simplified by using two SAS/IML matrix generating functions: (a) the J function, which creates a matrix having a specified number of rows and columns, with each element in the matrix equal to a single value, and (b) the I function, which produces an identity matrix of a given dimension.

By default, $\mathbf{U} = \mathbf{I}_1 = 1$; therefore, only the following statements are required to invoke the module for a test of the omnibus effect in this one-way design:

```
Y = {26, 34, . . . , 63, 59};
NX = {6  10  10};
C = J(2, 1, 1)||(−I(2));
PRINT 'TEST FOR OVERALL GROUP EFFECT';
RUN WJGLM;
```

In the first line, the commas used to separate the individual data points serve to delineate the rows of **Y**, so that **Y** is a column vector with 26 elements. In the third line, the J function is used to define a 2×1 vector of ones, and the negative I function specifies an identity matrix of dimension two. These two matrices are joined horizontally with the || operator to form

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}.$$

Alternatively, this **C** matrix could have been defined with the following SAS/IML (SAS Institute, 1989a) notation: C = {1 −1 0, 1 0 −1}.

The group means are 40.83, 48.00, and 57.50 for the low, moderate, and high conditions, respectively, with corresponding variances of 82.57, 150.22, and 95.83. The program output gives $T_{\mathrm{WJ}}/c = 5.84$, with $\nu_1 = 2$ and $\nu_2 = 13.78$ ($p = .0146$).

In performing all possible pairwise comparisons on the means, the RUN WJGLM statement is issued $J^* = J(J - 1)/2$ times, and **C** is redefined before each invocation. Accordingly, the additional programming lines for these results are as follows:

```
C = {1  −1  0};
PRINT 'CONTRAST J1 VS J2';
RUN WJGLM;
C = {1  0  −1};
PRINT 'CONTRAST J1 VS J3';
RUN WJGLM;
C = {0  1  −1};
PRINT 'CONTRAST J2 VS J3';
RUN WJGLM;
```

The program gives the following T_{WJ}/c, ν_2, and p value outputs, respectively, for the three comparisons: 1.78, 13.16, and .2042 (low–moderate); 11.90, 11.34, and .0052 (low–high); and 3.67, 17.16, and .0723 (moderate–high). For all comparisons, $\nu_1 = 1$.

Researchers conducting pairwise comparisons would typically wish to adopt a procedure for controlling the maximum familywise Type I error rate, that is, the probability of committing at least one Type I error when conducting multiple significance tests (Hayter, 1986). For illustration, we use a stepwise range procedure that combines the methods of Duncan (1957) and Shaffer (1979). The Duncan procedure was designed specifically for testing pairwise hypotheses on heteroscedastic means, whereas Shaffer suggested modifying the Studentized range critical values when a range procedure is used after a significant omnibus F test. This procedure can be used with Ryan (1960) and Welsch (1977) critical values to control the maximum familywise Type I error rate (Keselman & Lix, in press).

Because the omnibus test in the previous example was significant, one would step down to conduct pairwise comparisons. Only the low and high treatment means can be declared significantly different according to the combined Duncan (1957)-Shaffer (1979) method.

Factorial Between-Subjects Designs

Application of the general ADF solution for hypothesis testing in factorial between-subjects designs is discussed here only from the perspective of a two-way design. However, the same concepts may be readily extended to higher order designs.

Let Y_{ijk} represent the score associated with the ith subject in the (j, k)th treatment combination cell ($j = 1, \ldots, J$; $k = 1,$

$\ldots, K; i = 1, \ldots, n_{jk}; \Sigma_{j=1}^{J} \Sigma_{k=1}^{K} n_{jk} = N$). Adopting a cell means model,

$$Y_{ijk} = \mu_{jk} + \epsilon_{ijk}, \tag{9}$$

where μ_{jk} is the (j, k)th population mean and ϵ_{ijk} is the random error term. The Y_{ijk}s are assumed to be i.d. $N(\mu_{jk}, \sigma_{jk}^2)$ variates, with $\hat{\mu}_{jk}$ and $\hat{\sigma}_{jk}^2$, respectively, representing the (j, k)th sample mean and unbiased variance.

The sensitivity of the ANOVA F test to violations of its derivational assumptions for tests of main and interaction hypotheses in factorial designs has been studied in less detail than for one-way designs (Harwell et al., 1992). Nevertheless, the evidence available supports the conclusion that the F test may become seriously biased when equality of the σ_{jk}^2s is not a tenable assumption, particularly when the n_{jk}s are unequal (i.e., for nonorthogonal designs, in which hypotheses involving unweighted means are tested; Hsiung & Olejnik, 1994b; Milligan et al., 1987). According to Keselman et al. (in press), an ADF solution is largely robust in such situations.

For the simple 2×2 design, to test the interaction hypothesis H_{JK}: $\mu_{jk} - \mu_{j.} - \mu_{.k} + \mu_{..} = 0$ (for all j, k), where $\mu_{j.}$, $\mu_{.k}$, and $\mu_{..}$ are the respective population row (J), column (K), and grand means, the ADF statistic is

$$F(\mathbf{W}) = \frac{(\hat{\mu}_{11} - \hat{\mu}_{12} - \hat{\mu}_{21} + \hat{\mu}_{22})^2}{\sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\hat{\sigma}_{jk}^2}{n_{jk}}}. \tag{10}$$

For both main effects, the denominators of the test statistics are the same as for the interaction test of Equation 10. To test the J main effect (i.e., H_J: $\mu_{1.} = \mu_{2.}$), the numerator is $(\hat{\mu}_{1.} - \hat{\mu}_{2.})^2$; to test the K main effect (i.e., H_K: $\mu_{.1} = \mu_{.2}$), the numerator is $(\hat{\mu}_{.1} - \hat{\mu}_{.2})^2$. All three statistics approximately follow an F distribution with error degrees of freedom

$$\nu_{F(\mathbf{W})} = \frac{\left(\sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\hat{\sigma}_{jk}^2}{n_{jk}}\right)^2}{\sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(\hat{\sigma}_{jk}^2/n_{jk})^2}{n_{jk} - 1}}. \tag{11}$$

As Algina and Olejnik (1984) noted, testing main or interaction hypotheses when there are more than two levels of each factor, or more than two factors, is most easily accomplished by conceptualizing the test statistic in terms of matrix notation. With the GLM notation defined previously, for the two-factor design, $\mathbf{Y}$ is an $N \times 1$ vector with elements Y_{ijk}, $\mathbf{X}$ is the $N \times JK$ matrix with ones representing presence in a particular treatment combination cell and zeros representing absence, $\boldsymbol{\beta}^{\mathrm{T}} = (\mu_{11} \cdots \mu_{1K} \cdots \mu_{J1} \cdots \mu_{JK})$, and $\boldsymbol{\xi} = (\epsilon_{ijk})$.

To test the general linear hypothesis of Equation 2, $\mathbf{R} = \mathbf{C} = \mathbf{C}_{JK}$, $\mathbf{C}_J$, and $\mathbf{C}_K$, respectively, for tests of the interaction, row, and column hypotheses, because $\mathbf{U} = 1$ in all cases. Here, $\mathbf{C}_{JK} = \mathbf{C}_j \otimes \mathbf{C}_k$, where $\mathbf{C}_j$ and $\mathbf{C}_k$ are matrices of order $(J - 1) \times J$ and $(K - 1) \times K$, respectively, for which the rows represent sets of linearly independent contrasts among the levels of the between-subjects factors. Thus, $\mathbf{C}_{JK}$ is a contrast matrix of order $(J - 1)(K - 1) \times JK$. For the main effect tests, $\mathbf{C}_J = \mathbf{C}_j \otimes \mathbf{1}_k^{\mathrm{T}}$ and $\mathbf{C}_K = \mathbf{1}_j^{\mathrm{T}} \otimes \mathbf{C}_k$, where $\mathbf{1}_k$ and $\mathbf{1}_j$ are column vectors of ones (of order K and J, respectively) that serve to sum the means over the appropriate factor. Consequently, $\mathbf{C}_J$, a matrix of order $(J - 1) \times JK$, has $(J - 1)$ contrast rows that sum across the levels of factor K, and $\mathbf{C}_K$, a matrix of order $(K - 1) \times JK$, has $(K - 1)$ contrast rows that sum across the levels of factor J.

For higher order designs, Algina and Olejnik (1984) have developed a set of general rules that can be used to form $\mathbf{C}$. In an m factor design, in which letters are used to denote the factors, each $\mathbf{C}$ matrix can be represented as a Kronecker product of m matrices. For each letter that occurs as a subscript of $\mathbf{C}$, a contrast matrix of appropriate order occurs in the Kronecker product; for each letter that does not appear as a subscript of $\mathbf{C}$, a $\mathbf{1}^{\mathrm{T}}$ vector of appropriate order appears in the Kronecker product.

In nonorthogonal designs, researchers may test main effect hypotheses involving either weighted or unweighted means, depending on the values assigned to the elements of the $\mathbf{C}$ matrix (Maxwell & Delaney, 1990, pp. 94, 271–297). Keselman et al. (in press) have shown how different weighting schemes correspond to tests of different linear hypotheses and different model comparisons. For the sake of simplicity, we have assumed that the researcher is interested in testing hypotheses on unweighted means.[1] However, the specific strategy adopted should be a reflection of the hypotheses the researcher wishes to test.

Hsiung and Olejnik (1994a) have provided a nonpooled test statistic for contrasting pairs of either row or column marginal means. For example, the t statistic used to test $H_{kk'}$: $\mu_{.k} = \mu_{.k'}$ ($k \neq k'$) is

$$t(\mathbf{W}) = \frac{\sum_{j=1}^{J} (\hat{\mu}_{jk} - \hat{\mu}_{jk'})}{\sqrt{\sum_{j=1}^{J} \left(\frac{\hat{\sigma}_{jk}^2}{n_{jk}} + \frac{\hat{\sigma}_{jk'}^2}{n_{jk'}}\right)}}, \tag{12}$$

which approximately follows Student's t distribution with degrees of freedom

$$\nu_{t(\mathbf{W})} = \frac{\left[\sum_{j=1}^{J} \left(\frac{\hat{\sigma}_{jk}^2}{n_{jk}} + \frac{\hat{\sigma}_{jk'}^2}{n_{jk'}}\right)\right]^2}{\sum_{j=1}^{J} \left[\frac{(\hat{\sigma}_{jk}^2/n_{jk})^2}{n_{jk} - 1} + \frac{(\hat{\sigma}_{jk'}^2/n_{jk'})^2}{n_{jk'} - 1}\right]}. \tag{13}$$

For pairwise comparisons on the row marginal means using the general ADF solution, $\mathbf{R} = \mathbf{C} = \mathbf{c}_{jj'} \otimes \mathbf{1}_k^{\mathrm{T}}$, a $1 \times JK$ vector, where $\mathbf{c}_{jj'}$ contains the coefficients that contrast the jth and j'th row means. Similarly, when $\mathbf{R} = \mathbf{C} = \mathbf{1}_j^{\mathrm{T}} \otimes \mathbf{c}_{kk'}$, also a $1 \times JK$ vector, where $\mathbf{c}_{kk'}$ contains the coefficients that contrast the kth and k'th column means, a pairwise contrast on the column marginal means is formed.

A data set provided by Box and Cox (1964), and also considered by Algina and Olejnik (1984), was adopted to show how the SAS/IML (SAS Institute, 1989a) program for the general solution is used to test hypotheses in factorial designs. In a study of the effect of type of poison (factor J) and method of poison treatment (factor K) on animal survival rates, 4 animals were randomly assigned to each possible combination of poison

[1] Furthermore, a nonadditive model, in which none of the effects are constrained to zero, is assumed. To test main effect hypotheses assuming that certain effects are constrained, a restricted linear model must be adopted (Algina & Olejnik, 1984; Keselman et al., in press).

(three levels) and treatment (four levels). Although these data represent a balanced design, we selected them so as to be able to relate the results obtained with our program to those produced by other authors using a different computational algorithm.

The following SAS/IML (SAS Institute, 1989a) code produces the omnibus test results:

```
Y = {.31, .45, .46, .43, .82, , . . . , .30, .36, .31, .33};
NX = J(1, 12, 4);
CJ = J(2, 1, 1)‖(−I(2));
CK = J(3, 1, 1)‖(−I(3));
C = CJ@CK;
PRINT 'TEST FOR POISON × TREATMENT INTERACTION';
RUN WJGLM;
K4 = J(4, 1, 1);
C = CJ@K4';
PRINT 'TEST FOR POISON MAIN EFFECT';
RUN WJGLM;
J3 = J(3, 1, 1);
C = J3'@CK;
PRINT 'TEST FOR TREATMENT MAIN EFFECT';
RUN WJGLM;
```

Because cell sizes are equal, the J function in line 2 is used to define the **NX** vector, which contains 12 elements, each equal to 4. In lines 5, 9, and 13, the @ symbol is used to denote the Kronecker product of two matrices (vectors).[2] Also, in lines 9 and 13, the ' symbol represents the transpose operator.

Descriptive statistics for each Poison × Treatment combination were given by Algina and Olejnik (1984). The following results are obtained with the previous SAS/IML (SAS Institute, 1989a) code: For the test of the interaction, $T_{WJ}/c = 2.66$, $\nu_1 = 6$, and $\nu_2 = 10.55$ ($p = .0787$). The corresponding results for the poison main effect are 58.65, 2, and 10.68 ($p < .0001$); for the treatment main effect, they are 13.28, 3, and 8.58 ($p = .0014$).[3]

Because the interaction effect is not significant at $\alpha = .05$, but both main effects are significant, researchers may wish to probe the data using pairwise contrasts involving both row and column marginal means. The method for conducting these contrasts should be apparent from the previous example and the notation defined earlier in this section; therefore, only selected programming lines for comparing the J marginal means are provided. For example, in a comparison of the first and second levels of the poison factor (i.e., **C** = [1 1 1 1 −1 −1 −1 −1 0 0 0 0]), the code (assuming that the K4 vector was previously created for computing the omnibus test results) is

```
CJ12 = {1 −1 0};
C = CJ12@K4';
PRINT 'CONTRAST J1 VS J2';
RUN WJGLM;
```

The following T_{WJ}/c, ν_2, and p value outputs for the row marginal mean comparisons are obtained: J1–J2, 1.30, 10.58, and .2795; J1–J3, 104.26, 10.48, and <.0001; and J2–J3, 23.13, 6.51, and .0024. In all cases, $\nu_1 = 1$. Applying the combined Duncan (1957)-Shaffer (1979) method for control of the maximum familywise Type I error rate, only the comparison involving the J1 and J2 poison conditions cannot be declared significant.

One-Way Multivariate Between-Subjects Designs

In a one-way multivariate between-subjects experiment, n_j subjects ($\Sigma_{j=1}^{J} n_j = N$) in each of J groups are measured on a series of K dependent variables, and $\mathbf{Y}_{ij} = (Y_{ij1} \cdots Y_{ijK})$ represents the vector of scores associated with the ith subject in the jth group ($j = 1, \ldots, J$; $i = 1, \ldots, n_j$). The observations are modeled as

$$\mathbf{Y}_{ij} = \boldsymbol{\mu}_j + \boldsymbol{\epsilon}_{ij}, \tag{14}$$

where $\boldsymbol{\mu}_j = (\mu_{j1} \cdots \mu_{jK})$ and $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1} \cdots \epsilon_{ijK})$. The $\mathbf{Y}_{ij}$s are assumed to be i.d. $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ K-vector variates, with $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$, respectively, representing the jth sample mean vector and variance–covariance matrix.

When $J = 2$ and the data satisfy the assumptions of independence, normality, and covariance matrix equality, Hotelling's (1931) T^2 statistic provides the uniformly most powerful invariant and exact test of H_J: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. However, given that $\boldsymbol{\Sigma}_j$ has $K(K + 1)/2$ independent elements, it is unlikely that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. Both theoretical (Ito & Schull, 1964) and empirical (Algina & Oshima, 1990) studies provide evidence that, in the presence of covariance heterogeneity, Hotelling's T^2 produces biased results, particularly when the design is unbalanced.

When $J > 2$, one of four multivariate criteria may be adopted to test H_J: $\boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_J$, none of which is uniformly most powerful. These are the trace statistic of Pillai (1955) and Bartlett (1939), Roy's (1953) largest root criterion, Wilks's (1932) likelihood ratio, and the trace criterion of Hotelling (1951) and Lawley (1938). All four procedures are also sensitive to violations of the covariance homogeneity assumption, but to varying degrees (Olson, 1974; Stevens, 1979). Consequently, when the tenability of this assumption is in question, researchers may

---

[2] For this example, the linearly independent sets of contrast matrices that were created are as follows:

$$\mathbf{C} = \mathbf{C}_{JK} = \mathbf{C}_3 \otimes \mathbf{C}_4$$

$$= \begin{bmatrix} 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{C} = \mathbf{C}_J = \mathbf{C}_3 \otimes \mathbf{1}_4^{\mathrm{T}}$$

$$= \begin{bmatrix} 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \end{bmatrix},$$

and

$$\mathbf{C} = \mathbf{C}_K = \mathbf{1}_3^{\mathrm{T}} \otimes \mathbf{C}_4$$

$$= \begin{bmatrix} 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 \end{bmatrix}.$$

[3] Although Algina and Olejnik (1984) gave their solution in terms of chi-square rather than F statistics, their numerical results are duplicated with respect to the computed values of T_{WJ}, A, and ν_1.

choose to adopt an ADF solution, in this case Johansen's (1980) multivariate generalization of Welch's (1951) statistic.

In the two-group case, the general statistic reduces to (Algina et al., 1991)

$$T_{WJ} = \hat{\mu}_d \hat{\Sigma}_d^{-1} \hat{\mu}_d^{T}, \quad (15)$$

where $\hat{\mu}_d = (\hat{\mu}_1 - \hat{\mu}_2)$ and $\hat{\Sigma}_d = (\hat{\Sigma}_1/n_1 + \hat{\Sigma}_2/n_2)$. The statistic T_{WJ}/c is approximately distributed as an F variate and is referred to the critical value, $F[(1-\alpha); K, \nu_2]$, with both c and ν_2 as defined for the general solution.[4]

Generalizing this result, let $\mathbf{Y} = (\mathbf{Y}_{ij})$, the $N \times K$ matrix of subject scores. The design matrix, $\mathbf{X}$, has the same representation as for the one-way univariate design, $\boldsymbol{\beta}^{T} = (\boldsymbol{\mu}_1^{T} \cdots \boldsymbol{\mu}_J^{T})$, and $\boldsymbol{\xi} = (\epsilon_{ij})$. With respect to the linear hypothesis of Equation 2, $\mathbf{R} = \mathbf{C}_j \otimes \mathbf{I}_k$ because $\mathbf{C} = \mathbf{C}_j$, where $\mathbf{C}_j$ has the same form and function as for the univariate case, and $\mathbf{U} = \mathbf{U}^{T} = \mathbf{I}_k$. Accordingly, $\mathbf{R}$, which is of order $(J-1)K \times JK$, forms a set of linearly independent contrasts among the levels of the between-subjects factor for each dependent variable. In Equation 3, $\hat{\boldsymbol{\mu}} = (\hat{\mu}_{11} \cdots \hat{\mu}_{1K} \cdots \hat{\mu}_{J1} \cdots \hat{\mu}_{JK})^{T}$ and $\hat{\boldsymbol{\Sigma}} = \text{diag}(\hat{\boldsymbol{\Sigma}}_1/n_1 \cdots \hat{\boldsymbol{\Sigma}}_J/n_J)$.

In the one-way multivariate design, as in the univariate design, researchers may wish to test pairwise comparison hypotheses of the form $H_{jj'}$: $\mu_j = \mu_{j'} (j \neq j')$. Hotelling's (1931) T^2 represents the uniformly most powerful invariant and exact statistic for conducting these contrasts if equality of the group covariance matrices holds, along with other distributional assumptions. However, if it is unlikely that this assumption will be satisfied, the multivariate generalization of Welch's (1938) statistic should be adopted. Thus, to test pairwise hypotheses using the general ADF solution, $\mathbf{C} = \mathbf{c}_{jj'}$, where $\mathbf{c}_{jj'}$ has the same form and function as for the univariate design, and $\mathbf{U} = \mathbf{U}^{T} = \mathbf{I}_k$ so that $\mathbf{R} = \mathbf{c}_{jj'} \otimes \mathbf{I}_k$, a $K \times JK$ matrix.

To demonstrate the SAS/IML (SAS Institute, 1989a) program for testing omnibus and pairwise contrast hypotheses in a one-way multivariate between-subjects design, we have adopted an example data set from Stevens (1992, p. 231) for a three-group design in which the effects of an anxiety and social skills training program on female college students' responses in heterosexual encounters were studied. Participants were assigned to a control (C), behavioral rehearsal (BR), or behavioral rehearsal and cognitive restructuring (BR + CR) group. At the end of the training program, they were assessed on the dependent variables of physiological anxiety, social interaction skills, appropriateness of behavior, and assertiveness. Although 11 participants were assigned to each group, the responses associated with the last 4 participants in the third group have been eliminated to demonstrate how the program is applied to unbalanced designs.

Only the following lines of the SAS/IML (SAS Institute, 1989a) code are required to obtain a test of the group effect:

```
Y = {5 3 3 3, 5 4 4 3, . . . . , 4 4 4 4, 4 5 4 4};
NX = {11 11 7};
C = J(2, 1, 1)||(-I(2));
PRINT 'TEST FOR OVERALL GROUP EFFECT';
RUN WJGLM;
```

Commas are used to separate the data points associated with each participant and to define the rows of the data matrix, such that $\mathbf{Y}$ is of dimension 29×4. By default, the program sets $\mathbf{U} = \mathbf{I}_4$ because four scores are associated with each participant.[5] Using the previous statements, the program output gives $T_{WJ}/c = 4.54$, with $\nu_1 = 8$ and $\nu_2 = 15.71$ ($p = .0051$).

The code that produces pairwise contrasts on the group mean vectors is equivalent to the code for the one-way univariate between-subjects design and therefore is not repeated. For the contrast of the C and BR conditions, $T_{WJ}/c = 7.53$ ($\nu_2 = 14.75$, $p = .0016$); for the C and BR + CR comparison, $T_{WJ}/c = 0.9442$ ($\nu_2 = 11.19$, $p = .4739$); and, for the comparison of the BR and BR + CR conditions, $T_{WJ}/c = 8.38$ ($\nu_2 = 10.79$, $p = .0025$). For all three contrasts, $\nu_1 = 4$.

In one-way between-subjects designs where $J = 3$, Levin, Serlin, and Seaman (1994) recommended Fisher's (1935) two-stage least significant difference procedure for pairwise comparisons. In the previous example, because the omnibus test for the group effect was significant at $\alpha = .05$, each of the two-group tests is evaluated at the same level of significance. Consequently, in the current example, only the comparison involving the C and BR + CR conditions cannot be declared significant.

Within-Subjects Designs

Consider the design in which n_j subjects ($\Sigma_{j=1}^{J} n_j = N$) in each of J groups are measured on a single dependent variable at K points in time or under each of K treatments. The observations $\mathbf{Y}_{ij} = (Y_{ij1} \cdots Y_{ijK})$ $(j = 1, \ldots, J; i = 1, \ldots, n_j)$ can be modeled, from a cell means perspective, in the same manner as for the one-way multivariate design (i.e., Equation 14) and are assumed to be i.d. $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ K-vector variates, with $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$ denoting the jth sample mean vector and variance–covariance matrix, respectively.

To test the within-subjects main effect (i.e., H_K: $\mu_{.1} = \cdots = \mu_{.K}$), either a univariate or multivariate (Hotelling's [1931] T^2) statistic may be adopted, provided that the requisite assumptions are satisfied. With respect to the univariate (or mixed model) approach, in addition to the multivariate normality and

---

[4] In the two-group multivariate design, the statistic A reduces to

$$A = \frac{1}{2} \sum_{j=1}^{2} [\text{tr}(\mathbf{I}_p - \mathbf{W}^{-1}\mathbf{W}_j)^2 + \{\text{tr}(\mathbf{I}_p - \mathbf{W}^{-1}\mathbf{W}_j)\}^2]/(n_j - 1),$$

where $\mathbf{W}_j = \hat{\boldsymbol{\Sigma}}_j^{-1}$, $\mathbf{W} = \mathbf{W}_1 + \mathbf{W}_2$, and tr is the trace operator.

[5] For this example, $\mathbf{R}$ has the following form:

$$\mathbf{R} = \mathbf{C}_2 \otimes \mathbf{I}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}.$$

Thus, the first between-subjects contrast row of $\mathbf{C}$ is distributed across the four dependent variables, and this set of contrasts is contained in the first four rows of $\mathbf{R}$. Similarly, the next four rows of $\mathbf{R}$ contain the set of contrasts obtained by distributing the second contrast row of $\mathbf{C}$ across the dependent variables.

independence assumptions, a set of orthonormalized contrasts among the levels of the repeated measures factor must exhibit a common variance, a requirement known as sphericity or circularity (Huynh & Feldt, 1970; Rouanet & Lepine, 1970). Furthermore, it is assumed that the covariance matrices of these orthonormalized variables are equal across all levels of the between-subjects factor(s).[6] Jointly these assumptions are referred to as multisample sphericity (Huynh, 1978). When multisample sphericity is not a tenable assumption, an adjusted degrees of freedom univariate test (Quintana & Maxwell, 1994) may be adopted. However, the multivariate approach rests on a less restrictive set of assumptions, namely equality of the covariance matrices, in addition to multivariate normality and independence of errors.

Either a univariate or multivariate procedure may also be used to test the within-subjects interaction effect (i.e., H_{JK}: $\mu_{jk} - \mu_{j.} - \mu_{.k} + \mu_{..} = 0$, for all j, k). Specifically, any one of the four multivariate criteria enumerated previously for the one-way multivariate between-subjects design may be adopted if the assumptions detailed earlier are satisfied.

Keselman, Lix, and Keselman (1994) recommended the multivariate approach over the univariate approach if the necessary assumptions are satisfied, because the former provides an exact test, whereas the adjusted degrees of freedom procedures are only approximate. However, the literature reveals that both approaches for testing within-subjects effects are sensitive to the presence of heterogeneous covariance matrices, particularly when the design is unbalanced (Belli, 1988; Keselman & Keselman, 1990). Under such conditions, Keselman et al. (1994) recommended adopting the ADF solution described by Johansen (1980). Finally, to test the omnibus hypothesis of equality of the group means (i.e., H_J: $\mu_{1.} = \cdots = \mu_{J.}$), Welch's (1951) test, as defined with Equation 3, should be favored over the usual ANOVA F test when the assumption of variance homogeneity is not tenable.

Adopting the GLM notation of Equation 1, the repeated measurements are represented in the same manner as for the one-way multivariate between-subjects design. To test the general linear hypothesis, both $\mathbf{C}$ and $\mathbf{U}$ are defined in terms of the effect to be tested to create the appropriate $\mathbf{R}$ contrast matrix. For the within-subjects interaction effect, $\mathbf{R} = \mathbf{C}_j \otimes \mathbf{U}_k^{\mathrm{T}}$ because $\mathbf{C} = \mathbf{C}_j$, where $\mathbf{C}_j$ has the same form and function as for the one-way univariate between-subjects design, and $\mathbf{U} = \mathbf{U}_k$, where $\mathbf{U}_k$ is a $K \times (K-1)$ matrix whose columns form a set of linearly independent contrasts among the levels of the within-subjects factor. Thus, $\mathbf{R}$ is of order $(J-1)(K-1) \times JK$. For tests of the within-subjects main effect, $\mathbf{C} = \mathbf{1}_j^{\mathrm{T}}$ and $\mathbf{U} = \mathbf{U}_k$; for the between-subjects main effect, $\mathbf{C} = \mathbf{C}_j$ and $\mathbf{U} = \mathbf{1}_k$. Consequently, these $\mathbf{R}$ matrices are of order $(K-1) \times JK$ and $(J-1) \times JK$, respectively.

A significant within-subjects interaction effect could be probed with a variety of procedures, including tetrad contrasts. These contrasts are used to test for the presence of an interaction in a 2×2 submatrix of the $J \times K$ data matrix (Boik, 1993; Timm, 1994). Lix and Keselman (in press) have demonstrated, however, that multiple comparison procedures based on a test statistic using a pooled estimate of error variance are not robust to departures from the multisample sphericity assumption. To test $H_{jj'kk'}$: $(\mu_{jk} - \mu_{jk'}) - (\mu_{j'k} - \mu_{j'k'}) = 0$, where $j \neq j'$ and $k \neq k'$, Lix and Keselman recommended a test statistic that uses a nonpooled estimate of error variance

$$t(\mathrm{W}) = \frac{(\hat{\mu}_{jk} - \hat{\mu}_{jk'}) - (\hat{\mu}_{j'k} - \hat{\mu}_{j'k'})}{\sqrt{\dfrac{\mathbf{d}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_j\mathbf{d}}{n_j} + \dfrac{\mathbf{d}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{j'}\mathbf{d}}{n_{j'}}}}, \tag{16}$$

where $\mathbf{d}$ is a $K \times 1$ vector of coefficients that contrasts the kth and k'th means. This statistic approximately follows a t distribution with degrees of freedom

$$\nu_{t(\mathrm{W})} = \frac{\left(\dfrac{\mathbf{d}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_j\mathbf{d}}{n_j} + \dfrac{\mathbf{d}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{j'}\mathbf{d}}{n_{j'}}\right)^2}{\dfrac{(\mathbf{d}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_j\mathbf{d}/n_j)^2}{n_j - 1} + \dfrac{(\mathbf{d}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{j'}\mathbf{d}/n_{j'})^2}{n_{j'} - 1}}. \tag{17}$$

Tetrad contrast results are also obtainable with the general ADF solution, eliminating the need to rely on the specific formulas of Equations 16 and 17. Here, $\mathbf{R} = \mathbf{c}_{jj'} \otimes \mathbf{u}_{kk'}^{\mathrm{T}}$ because $\mathbf{C} = \mathbf{c}_{jj'}$ and $\mathbf{U} = \mathbf{u}_{kk'}$, where $\mathbf{u}_{kk'}$ is a column vector of coefficients (i.e., u_ks) that contrasts the kth and k'th within-subjects means ($\Sigma_{k=1}^{K} u_k = 0$). For such contrasts, $\mathbf{R}$ is of order $1 \times JK$.

The within-subjects main effect may be probed with pairwise comparisons of the marginal means. However, multiple comparison procedures that rely on a statistic based on the usual pooled estimate of error variance are known to be sensitive to violations of the multisample sphericity assumption (Keselman & Keselman, 1988). Keselman et al. (1991) found that tests of pairwise comparison hypotheses that are largely insensitive to departures from this assumption may be obtained by adopting a t statistic that uses a nonpooled estimate of error variance (see also Keselman, 1994). To test $H_{kk'}$: $\mu_{.k} = \mu_{.k'} (k \neq k')$ with an unweighted means solution, the nonpooled statistic is

$$t(\mathrm{W}) = \frac{\sum_{j=1}^{J} (\hat{\mu}_{jk} - \hat{\mu}_{jk'})/J}{\sqrt{\dfrac{1}{J^2}\sum_{j=1}^{J} \dfrac{\mathbf{d}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_j\mathbf{d}}{n_j}}}, \tag{18}$$

where $\mathbf{d}$ is again a $K \times 1$ vector of coefficients that contrasts the kth and k'th means. This statistic is approximately distributed as Student's t with degrees of freedom

$$\nu_{t(\mathrm{W})} = \frac{\left(\dfrac{1}{J^2}\sum_{j=1}^{J} \dfrac{\mathbf{d}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_j\mathbf{d}}{n_j}\right)^2}{\sum_{j=1}^{J} \dfrac{(\mathbf{d}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_j\mathbf{d})^2/n_j^2}{J^4(n_j - 1)}}. \tag{19}$$

To test these pairwise comparison hypotheses using a GLM ap-

[6] Neither single factor nor multifactor within-subjects designs are considered because covariance heterogeneity is not an issue when the design does not contain a between-subjects grouping variable. Thus, the T_{WJ} statistic is not an option. However, one should always use a procedure and critical value that can either contend with violations of the sphericity assumption (e.g., an adjusted degrees of freedom test; Quintana & Maxwell, 1994) or bypass the assumption altogether (e.g., a multivariate test).

proach, $\mathbf{R} = \mathbf{1}_J^T \otimes \mathbf{u}_{kk'}^T$ and is of order $1 \times JK$, because $\mathbf{C} = \mathbf{1}_J^T$ and $\mathbf{U} = \mathbf{u}_{kk'}$.

Application of the SAS/IML (SAS Institute, 1989a) program for the general ADF solution to a within-subjects design is illustrated with an example data set from Maxwell and Delaney (1990, pp. 518–523). In this study, participants were classified into two groups on the basis of age and were required to detect a target letter displayed at each of three different viewing angles on a screen, with the dependent variable being reaction time performance. Although 10 participants were originally randomly assigned to each group, the scores associated with the last 3 participants in the first group have been deleted to create an unbalanced design.

Summary statistics for the analysis were provided by Keselman et al. (1993), because the same data set was analyzed with an SAS/IML (SAS Institute, 1989a) program written for this specific ADF application. The following programming lines produce the omnibus test results for the general solution:

```
Y = {450 510 630, 390 480 540, . . . , 510 690 810};
NX = {7  10};
C1 = J(1, 1, 1)‖(−I(1));
C = C1;
U = J(1, 2, 1)//(−I(2));
PRINT 'TEST FOR AGE × ANGLE INTERACTION';
RUN WJGLM;
C2 = J(2, 1, 1);
C = C2';
PRINT 'TEST OF ANGLE MAIN EFFECT';
RUN WJGLM;
C = C1;
U = J(3, 1, 1);
PRINT 'TEST OF AGE MAIN EFFECT';
RUN WJGLM;
```

In line 5, the J function is used to define a 1×2 vector of ones, and the negative I function specifies an identity matrix of dimension two. These two matrices are joined vertically with the // operator to form the U matrix, such that the columns of U form a set of linearly independent contrasts among the levels of the within-subjects factor. Because the same U matrix is used to test both the within-subjects main effect and the interaction effect, it need not be respecified before the second invocation of the module.

For the interaction effect, $T_{WJ}/c = 6.44$, $\nu_1 = 2$, and $\nu_2 = 10.00$ ($p = .0159$); for the within-subjects main effect, $T_{WJ}/c = 81.88$ with 2 and 10.00 df ($p < .0001$). Because there are only two between-subjects factor levels, the test for an age main effect is equivalent to Welch's (1938) t test and the ADF test statistic equals 7.83 ($\nu_1 = 1$, $\nu_2 = 13.30$, $p = .0147$).

Keselman et al. (1993) suggested that, to obtain a robust test of the within-subjects interaction effect hypothesis using an ADF solution, the ratio of the smallest n_j to $(K - 1)$ should be at least 3 or 4:1, and preferably higher if the validity of the multivariate normality assumption is questionable. When there is an insufficient number of subjects to meet this requirement, the authors suggested adopting a .01 criterion of significance to maintain the error rate below 5%. Consequently, in the current example, the within-subjects interaction effect cannot be declared significant.

This nonsignificant result would preclude further probing of the interaction; however, the additional SAS/IML code lines required to produce interaction contrast results are given for demonstration purposes:

```
C = C1;
U12 = {1, −1, 0};
U = U12;
PRINT 'INTERACTION CONTRAST WITH K1 AND K2';
RUN WJGLM;
U13 = {1, 0, −1};
U = U13;
PRINT 'INTERACTION CONTRAST WITH K1 AND K3';
RUN WJGLM;
U23 = {0, 1, −1};
U = U23;
PRINT 'INTERACTION CONTRAST WITH K2 AND K3';
RUN WJGLM;
```

In lines 2, 6, and 10, commas are used to separate the rows of the U matrix so that, in each instance, a column of contrast coefficients is created. Furthermore, because there are only two levels of the between-subjects grouping factor, the same C matrix is used to define all interaction contrasts; for $J > 2$, however, this would not be the case.[7]

The results obtained from the program are as follows: For the tetrad contrast involving K1 and K2, $T_{WJ}/c = 0.02$ ($\nu_2 = 12.85$, $p = .8866$); for the K1 and K3 contrast, $T_{WJ}/c = 8.59$ ($\nu_2 = 11.65$, $p = .0129$); and, for the contrast involving K2 and K3, $T_{WJ}/c = 12.45$ ($\nu_2 = 9.72$, $p = .0057$). In all cases, the program gives $\nu_1 = 1$.

Lix and Keselman (in press) found that control of the maximum familywise Type I error rate for all possible tetrad contrasts could be obtained with Shaffer's (1986) modified sequentially rejective Bonferroni procedure. Applying this procedure to the set of J^*K^* hypotheses considered here, where $K^* = K(K - 1)/2$, only the tetrad contrast involving the second and third levels of the within-subjects factor is significant.

The following SAS/IML (SAS Institute, 1989a) code is used to test all possible pairs of repeated measures marginal means:

```
C = C2';
U = U12;
PRINT 'CONTRAST K1 VS K2';
RUN WJGLM;
U = U13;
PRINT 'CONTRAST K1 VS K3';
RUN WJGLM;
U = U23;
PRINT 'CONTRAST K2 VS K3';
RUN WJGLM;
```

For the contrast of K1 and K2, $T_{WJ}/c = 67.65$ ($\nu_2 = 12.85$, $p <$

[7] If there were three levels of the between-subjects factor, and the researcher were interested in testing the interaction hypothesis involving J1 and J3 and K1 and K3, the following programming lines would be used:

```
C = {1  0  −1};
U = {1, −1, 0};
PRINT 'INTERACTION CONTRAST WITH J1, J3, K1, & K2';
RUN WJGLM;
```

.0001); for the K1 and K3 contrast, the computed value is 174.56 ($\nu_2 = 11.65$, $p < .0001$); and, for the K2 and K3 contrast, the value is 59.49 ($\nu_2 = 9.72$, $p < .0001$). Again, $\nu_1 = 1$ for all contrasts. Applying the combined Duncan (1957)-Shaffer (1979) method to the current example (see also Keselman & Lix, in press), all three comparisons are significant at $\alpha = .05$.

A New Application of the General ADF Test Statistic

The previous section illustrated how specific formulas for ADF solutions, as presented in the literature, may be replaced by the general formulation of Johansen (1980). This general solution may also be applied to research designs that have not previously been considered with respect to ADF tests. One example is multivariate within-subjects designs (Boik, 1991; Timm, 1980, pp. 68–74).

In the simplest design containing a single between-subjects factor, n_j subjects ($\Sigma_{j=1}^{J} n_j = N$) in each of J groups are measured on each of L dependent variables over K occasions or trials. Thus, $\mathbf{Y}_{ij} = (Y_{ij11} \cdots Y_{ij1L} \cdots Y_{ijK1} \cdots Y_{ijKL})$ represents the vector of KL scores associated with the ith subject in the jth group, and

$$\mathbf{Y}_{ij} = \boldsymbol{\mu}_j + \boldsymbol{\epsilon}_{ij}, \tag{20}$$

where $\boldsymbol{\mu}_j = (\mu_{j11} \cdots \mu_{j1L} \cdots \mu_{jK1} \cdots \mu_{jKL})$ and $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij11} \cdots \epsilon_{ij1L} \cdots \epsilon_{ijK1} \cdots \epsilon_{ijKL})$. The $\mathbf{Y}_{ij}$s are assumed to be i.d. $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ KL-vector variates with $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$, respectively, representing the jth sample mean vector and variance–covariance matrix.

Multivariate repeated measures data may be modeled from either a multivariate mixed model or a doubly multivariate perspective. With both approaches, a multivariate criterion, such as the trace statistic of Pillai (1955) and Bartlett (1939), is used to test the within-subjects main and interaction effects. However, each approach rests on its own set of derivational assumptions. No restrictions are placed on the structure of the pooled covariance matrix under the doubly multivariate approach. However, the assumptions of independence, multivariate normality, and homogeneity of the group orthonormalized covariance matrices for some linear combination of the dependent variables must hold. In contrast, the multivariate mixed model approach assumes that the multivariate multisample sphericity assumption is satisfied (Boik, 1991), in addition to the multivariate normality and independence assumptions. Finally, a test of the multivariate between-subjects effect is obtained in the same manner as for the one-way multivariate design if the requisite assumptions are met. Because behavioral science data are unlikely to satisfy the assumptions relating to covariance homogeneity, it may be more appropriate to adopt ADF statistics to test each of these multivariate effects.

Returning to the GLM of Equation 1, $\mathbf{Y} = (\mathbf{Y}_{ij})$, the $N \times KL$ matrix of subject scores, where the first L columns of $\mathbf{Y}$ correspond to the dependent variable scores obtained at the first level of the repeated measures factor, the next L columns correspond to the dependent variable scores for the second repeated measurement, and so on. The design matrix, $\mathbf{X}$, has the same structure as for the one-way between-subjects design with a single dependent variable, $\boldsymbol{\beta}^{\mathsf{T}} = (\boldsymbol{\mu}_1^{\mathsf{T}} \cdots \boldsymbol{\mu}_J^{\mathsf{T}})$, and $\boldsymbol{\xi} = (\epsilon_{ij})$.

Forming the $\mathbf{R}$ matrix to test the general linear hypothesis is more complex than for other designs but is easily accomplished with the notation developed in the previous section. The structure of $\mathbf{R}$ will depend on the effect to be tested. Let $\boldsymbol{\mu}_{jk}$ denote the $1 \times L$ mean vector for the jth group on the kth trial. The multivariate within-subjects interaction hypothesis is H_{JK}:$\boldsymbol{\mu}_{jk} - \boldsymbol{\mu}_{j.} - \boldsymbol{\mu}_{.k} + \boldsymbol{\mu}_{..} = \mathbf{0}$ (for all j, k), where $\boldsymbol{\mu}_{j.}$, $\boldsymbol{\mu}_{.k}$, and $\boldsymbol{\mu}_{..}$, respectively, represent the unweighted jth group, kth repeated measure, and grand mean vectors. Using the notation of Equation 2 to specify this hypothesis, $\mathbf{R} = \mathbf{C}_j \otimes (\mathbf{U}_k \otimes \mathbf{I}_l)^{\mathsf{T}}$ and is of order $(J-1)(K-1)L \times JKL$, because $\mathbf{C} = \mathbf{C}_j$ and $\mathbf{U} = (\mathbf{U}_k \otimes \mathbf{I}_l)$. As a result of specifying this form for $\mathbf{U}$, the first within-subjects contrast column of $\mathbf{U}_k$ is distributed across the L dependent variables, and this set of contrasts is contained in the first L columns of $\mathbf{U}$. Similarly, the next L columns of $\mathbf{U}$ contain the set of contrasts obtained by distributing the second contrast column of $\mathbf{U}_k$ across the dependent variables, and so on. To test the within-subjects main effect hypothesis, H_K: $\boldsymbol{\mu}_{.1} = \cdots = \boldsymbol{\mu}_{.K}$, $\mathbf{R} = \mathbf{1}_j^{\mathsf{T}} \otimes (\mathbf{U}_k \otimes \mathbf{I}_l)^{\mathsf{T}}$ because $\mathbf{C} = \mathbf{1}_j^{\mathsf{T}}$. Here, $\mathbf{R}$ is of order $(K-1)L \times JKL$. The $\mathbf{U}$ contrast matrix for this test is equivalent to the one used to test the within-subjects interaction effect. Finally, when $\mathbf{R}$ represents the matrix product involving $\mathbf{C} = \mathbf{C}_j$ and $\mathbf{U} = (\mathbf{1}_k \otimes \mathbf{I}_l)$, a test of the between-subjects main effect hypothesis H_J: $\boldsymbol{\mu}_{1.} = \cdots = \boldsymbol{\mu}_{J.}$ is achieved. In this case, $\mathbf{R}$ is of order $(J-1)L \times JKL$. In Equation 3, $\hat{\boldsymbol{\mu}} = (\hat{\mu}_{111} \cdots \hat{\mu}_{11L} \cdots \hat{\mu}_{JK1} \cdots \hat{\mu}_{JKL})^{\mathsf{T}}$ and $\hat{\boldsymbol{\Sigma}} = \text{diag}(\hat{\boldsymbol{\Sigma}}_1/n_1 \cdots \hat{\boldsymbol{\Sigma}}_J/n_J)$, where $\hat{\boldsymbol{\Sigma}}_j$ is of order KL.

Data for a multivariate within-subjects design provided by Rich (1983) are used here to demonstrate the SAS/IML (SAS Institute, 1989a) program for testing within-subjects and between-subjects omnibus effects with the general ADF solution. The data are from a study examining mood pattern changes over time among introverts and extraverts. Ten introverts and 9 extraverts were measured on the variables of elation and depression over the course of 5 days. The number of observations may be too small to provide a robust test, but these data are used for purposes of illustration only.

The code for testing the omnibus effects is as follows:

```
Y = {20 30 21 35 42 20 32 27 32 22, , . . . , 64 34 53 61
62 36 40 54 43 47};
NX = {10  9};
C1 = J(1, 1, 1)‖(−I(1));
C = C1;
U1 = J(1, 4, 1)//(−I(4));
I2 = I(2);
U = U1@I2;
PRINT 'TEST OF SOCIAL GROUP BY DAYS INTERACTION';
RUN WJGLM;
C2 = J(2, 1, 1);
C = C2';
PRINT 'TEST OF DAYS MAIN EFFECT';
RUN WJGLM;
C = C1;
U1 = J(5, 1, 1);
U = U1@I2;
PRINT 'TEST OF SOCIAL GROUP MAIN EFFECT';
RUN WJGLM;
```

In line 1, commas are again used to delineate the rows of $\mathbf{Y}$, and, in a particular row, elation and depression scores for the

1st day are followed by the two scores for the 2nd day, and so on.[8]

For the multivariate within-subjects interaction effect, $T_{WJ}/c = 0.40$, with $\nu_1 = 8$ and $\nu_2 = 11.91$ ($p = .8981$); for the within-subjects main effect, $T_{WJ}/c = 17.59$, with 8 and 11.90 df ($p < .0001$). Finally, for the between-subjects main effect, the test statistic equals 5.19, with 2 and 13.94 df ($p = .0206$).

In this analysis, tests of the omnibus within-subjects and between-subjects effects were applied to both dependent variables simultaneously. Subsequent analyses might include multiple multivariate interaction or marginal mean contrasts or tests of omnibus effects conducted separately for each dependent variable. Regardless of the analysis strategy adopted, ADF test statistics may be computed with the general formulation of Equation 3.

Summary

In this article, we have presented one statistic, T_{WJ}, that can be used to obtain an approximate degrees of freedom test for mean equality in between-subjects and within-subjects designs. As we have demonstrated, this statistic can be used to test main and interaction effects, as well as multiple comparison hypotheses related to these effects. Thus, researchers need only be familiar with this statistic, rather than the many statistics that have been defined in the literature, to perform these tests of significance. We also included a program that enables researchers to obtain a numerical solution by providing a few lines of simple input. Finally, we indicated how this statistic can be used to arrive at tests of significance in other problems not previously discussed in the literature by demonstrating its use for testing repeated measures effects when there are multiple dependent variables associated with each repeated measurement.

In conclusion, we believe that it is important to reiterate that an ADF statistic is not a panacea for investigating mean differences in all cases. The empirical literature indicates that, like classical tests for mean equality (see Sawilowsky & Blair, 1992), the ADF procedure is affected by extreme skewness of the underlying distribution (e.g., Algina et al., 1991; Keselman et al., 1993, in press). Nonetheless, the literature also suggests that, for many types and degrees of nonnormality likely to be encountered by behavioral scientists (e.g., see Keselman et al., 1993, in press; Micceri, 1989; Sawilowsky & Blair, 1992), the T_{WJ} statistic should perform quite well with regard to Type I error control. In addition, the literature shows that the ADF approach compares favorably with the classical approach in terms of statistical power.

[8] It is important to note that the data could have been entered in a different order such that, in a particular row, elation scores for all days are followed by depression scores for all days. If this order of data entry were used, $\mathbf{U} = (\mathbf{I}_t \otimes \mathbf{U}_k)$, and in lines 8 and 17 of the program, the code U= I2@U1 would have to be used to obtain the correct contrast matrix.

References

Algina, J., & Olejnik, S. F. (1984). Implementing the Welch-James procedure with factorial designs. *Educational and Psychological Measurement, 44,* 39–48.

Algina, J., & Oshima, T. C. (1990). Robustness of the independent samples Hotelling's T^2 to variance-covariance heteroscedasticity when sample sizes are unequal and in small ratios. *Psychological Bulletin, 108,* 308–313.

Algina, J., Oshima, T. C., & Tang, K. L. (1991). Robustness of Yao's, James', and Johansen's tests under variance-covariance heteroscedasticity and nonnormality. *Journal of Educational Statistics, 16,* 125–139.

Bartlett, M. S. (1939). A note on tests of significance in multivariate analysis. *Proceedings of the Cambridge Philosophical Society, 35,* 180–185.

Behrens, W. V. (1929). A contribution to the calculation of error for few observations. *Landwirtsch Jahrbücher, 68,* 807–837.

Belli, G. M. (1988, April). *Type I error rates of MANOVA of repeated measures under group heterogeneity in unbalanced designs.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Boik, R. J. (1991). Scheffe's mixed model for multivariate repeated measures: A relative efficiency evaluation. *Communications in Statistics, Theory and Methods, 20,* 1233–1255.

Boik, R. J. (1993). The analysis of two-factor interactions in fixed effects linear models. *Journal of Educational Statistics, 18,* 1–40.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B, 26,* 211–243.

Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics, 16,* 129–132.

Chenier, T. C., & Seaman, S. L. (1987). A PASCAL program for implementing the Welch-James procedure. *Educational and Psychological Measurement, 47,* 123–125.

Dijkstra, J. B., & Werter, P. S. P. J. (1981). Testing the equality of several means when the population variances are unequal. *Communications in Statistics, Simulation and Computation, B10,* 557–569.

Dixon, W. J., Brown, M. B., Engelman, L., Hill, M. A., & Jennrich, R. I. (1988). *BMDP statistical software manual* (Vol. 1). Los Angeles: University of California Press.

Duncan, D. B. (1957). Multiple range tests for correlated and heteroscedastic means. *Biometrics, 13,* 164–176.

Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association, 75,* 796–800.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics, 6,* 391–398.

Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal *n*'s and/or variances: A Monte Carlo study. *Journal of Educational Statistics, 1,* 113–125.

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics, 17,* 315–339.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association, 81,* 1000–1004.

Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics, 2,* 360–378.

Hotelling, H. (1951). A generalized *t* test and measure of multivariate dispersion. In J. Neyman (Ed.), *Proceedings of the second Berkeley symposium on mathematical statistics and probability* (pp. 23–41). Berkeley: University of California Press.

Hsiung, T., & Olejnik, S. (1994a). Contrast analyses for additive nonorthogonal two-factor designs in unequal variance cases. *British Journal of Mathematical and Statistical Psychology, 47,* 337–354.

Hsiung, T., & Olejnik, S. (1994b, April). *Type I error rates and statisti-*

cal power for the James second-order test and the univariate F test in two-way fixed-effects ANOVA models under heteroscedasticity and/or nonnormality. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hsiung, T., Olejnik, S., & Oshima, T. C. (1994). A SAS/IML program for applying the James second-order test in two-factor fixed-effect ANOVA models. *Educational and Psychological Measurement, 54,* 696–698.

Huynh, H. (1978). Some approximate tests for repeated measurement designs. *Psychometrika, 43,* 161–175.

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,* 1582–1585.

Ito, K., & Schull, W. J. (1964). On the robustness of the T_0^2 test in multivariate analysis of variance when variance-covariance matrices are not equal. *Biometrika, 51,* 71–82.

James, G. S. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. *Biometrika, 41,* 19–43.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika, 67,* 85–92.

Keselman, H. J. (1994). Stepwise and simultaneous multiple comparison procedures of repeated measures' means. *Journal of Educational Statistics, 19,* 127–162.

Keselman, H. J., Carriere, K. C., & Lix, L. M. (1993). Testing repeated measures hypotheses when covariance matrices are heterogeneous. *Journal of Educational Statistics, 18,* 305–319.

Keselman, H. J., Carriere, K. C., & Lix, L. M. (in press). Robust and powerful nonorthogonal analyses. *Psychometrika.*

Keselman, H. J., & Keselman, J. C. (1990). Analysing unbalanced repeated measures designs. *British Journal of Mathematical and Statistical Psychology, 43,* 265–282.

Keselman, H. J., Keselman, J. C., & Shaffer, J. P. (1991). Multiple pairwise comparisons of repeated measures means under violation of multisample sphericity. *Psychological Bulletin, 110,* 162–170.

Keselman, H. J., & Lix, L. M. (in press). Improved repeated measures stepwise multiple comparison procedures. *Journal of Educational and Behavioral Statistics.*

Keselman, J. C., & Keselman, H. J. (1988). Repeated measures multiple comparison procedures: Effect of violating multisample sphericity in unbalanced designs. *Journal of Educational Statistics, 13,* 215–226.

Keselman, J. C., Lix, L. M., & Keselman, H. J. (1994, April). *The analysis of repeated measurements: A quantitative research synthesis.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole.

Lawley, D. N. (1938). A generalization of Fisher's *z* test. *Biometrika, 30,* 180–187.

Levin, J. R., Serlin, R. C., & Seaman, M. A. (1994). A controlled, powerful multiple-comparison strategy for several situations. *Psychological Bulletin, 115,* 153–159.

Lix, L. M., & Keselman, H. J. (1994). [To trim or not to trim: Tests of mean equality under heteroscedasticity and nonnormality]. Unpublished raw data.

Lix, L. M., & Keselman, H. J. (in press). Interaction contrasts in repeated measures designs. *British Journal of Mathematical and Statistical Psychology.*

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

Micceri, T. (1989). The unicorn, the normal distribution, and other improbable creatures. *Psychological Bulletin, 105,* 156–166.

Milligan, G. W., Wong, D. W., & Thompson, P. A. (1987). Robustness properties of nonorthogonal analysis of variance. *Psychological Bulletin, 101,* 464–470.

Norusis, M. J. (1990). *SPSS introductory statistics.* New York: McGraw-Hill.

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association, 69,* 894–908.

Oshima, T. C., & Algina, J. (1992). A SAS program for testing the hypothesis of equal means under heteroscedasticity: James's second-order test. *Educational and Psychological Measurement, 52,* 117–118.

Pillai, K. C. S. (1955). Some new test criteria in multivariate analysis. *Annals of Mathematical Statistics, 26,* 117–121.

Quintana, S. M., & Maxwell, S. E. (1994). A Monte Carlo comparison of seven ϵ-adjustment procedures in repeated measures designs with small sample sizes. *Journal of Educational Statistics, 19,* 57–72.

Rich, C. E. (1983). Repeated measures designs. In R. W. Barcikowski (Ed.), *Computer packages and research design: With annotations of input and output from the BMDP, SAS, and SPSSX statistical packages, Vol. 3: SPSS and SPSSX* (pp. 567–710). Lanham, MD: University Press of America.

Roth, A. J. (1983). Robust trend tests derived and simulated: Analogs of the Welch and Brown-Forsythe tests. *Journal of the American Statistical Association, 78,* 972–980.

Roth, A. J. (1988). Welch tests. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 9, pp. 586–589). New York: Wiley.

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

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics, 24,* 220–238.

Ryan, T. A. (1960). Significance tests for multiple comparisons of proportions, variances and other statistics. *Psychological Bulletin, 57,* 318–328.

SAS Institute. (1989a). *SAS/IML software: Usage and reference, Version 6.* Cary, NC: Author.

SAS Institute. (1989b). *SAS/STAT user's guide, Version 6* (4th ed., Vol. 2). Cary, NC: Author.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research, 60,* 91–126.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error probabilities of the *t* test to departures from population normality. *Psychological Bulletin, 111,* 352–360.

Shaffer, J. P. (1979). Comparison of means: An *F* test followed by a modified multiple range procedure. *Journal of Educational Statistics, 4,* 14–23.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association, 81,* 826–831.

Stevens, J. (1979). Comment on Olson: Choosing a test statistic in multivariate analysis. *Psychological Bulletin, 86,* 355–360.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Erlbaum.

Tang, K. L., & Algina, J. (1993). Performance of four multivariate tests under variance-covariance heteroscedasticity. *Multivariate Behavioral Research, 28,* 391–405.

Timm, N. H. (1975). *Multivariate analysis with applications in education and psychology.* Monterey, CA: Brooks/Cole.

Timm, N. H. (1980). Multivariate analysis of variance of repeated measurements. In P. R. Krishnaiah (Ed.), *Handbook of statistics, Vol. 1: Analysis of variance* (pp. 41–88). Amsterdam: North-Holland.

Timm, N. H. (1994, April). *Analysis of interactions: Another look.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin, 99,* 90–99.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29,* 350–362.

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika, 34,* 28–35.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38,* 330–336.

Welsch, R. E. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association, 72,* 566–575.

Wilcox, R. R. (1988). A new alternative to the ANOVA F and new results on James's second-order method. *British Journal of Mathematical and Statistical Psychology, 41,* 109–117.

Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W, and F^* statistics. *Communications in Statistics, Simulation and Computation, 15,* 933–943.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika, 24,* 471–494.

Appendix A

Details of the ADF Solution

From a GLM perspective, the matrix of population means, $\boldsymbol{\beta}$, is estimated by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Let $\mathbf{Y}_j = \mathbf{Y} \cdot (\mathbf{X}_j\mathbf{1}_p^T)$ be a Hadamard product, where $\mathbf{X}_j$ is the jth column of $\mathbf{X}$ ($j = 1, \ldots, r$) and consists entirely of zeros and ones, $\mathbf{1}_p$ is a $p \times 1$ vector of ones, superscript T is the transpose operator, and "$\cdot$" is the dot product function, such that $\mathbf{Y}_j$ is an element-by-element product matrix. Then

$$\hat{\boldsymbol{\Sigma}}_j = \frac{(\mathbf{Y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}_j)^T(\mathbf{Y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}_j)}{n_j - 1}$$

provides an estimate of $\boldsymbol{\Sigma}_j$.

The statistic A is given by

$$A = \frac{1}{2}\sum_{j=1}^{r}[\mathrm{tr}\{\hat{\boldsymbol{\Sigma}}\mathbf{R}^T(\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T)^{-1}\mathbf{R}\mathbf{Q}_j\}^2 + \{\mathrm{tr}(\hat{\boldsymbol{\Sigma}}\mathbf{R}^T(\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T)^{-1}\mathbf{R}\mathbf{Q}_j)\}^2]/(n_j - 1),$$

where tr is the trace operator. In a multivariate model, $\mathbf{Q}_j$ is a symmetric block matrix of dimension $r \times p$ associated with $\mathbf{X}_j$, such that the (s, t)th diagonal block of $\mathbf{Q}_j = \mathbf{I}_p$ if $s = t = j$ and is $\mathbf{0}$ otherwise.

In univariate designs, that is, where $p = 1$,

$$\hat{\sigma}_j^2 = \frac{(\mathbf{Y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}_j)^T(\mathbf{Y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}_j)}{n_j - 1}$$

provides an estimate of σ_j^2. In computing the statistic A, $\mathbf{Q}_j$ is a diagonal matrix of dimension r such that the (s, t)th element of $\mathbf{Q}_j$ is 1 if $s = t = j$ and is 0 otherwise.

Appendix B

SAS/IML Program for the ADF Solution

```
***INVOKE THE IML PROCEDURE & DEFINE THE MODULE WJGLM***;
PROC IML;
RESET NONAME;
START WJGLM;
**********PERFORM DIAGNOSTICS AND DEFINE MATRICES**********;
IF NROW(U) = 0 THEN U = I(NCOL(Y));
IF NROW(C) > NCOL(C) THEN PRINT
  'ERROR: NUMBER OF ROWS OF C EXCEEDS NUMBER OF COLUMNS';
IF NCOL(U) > NROW(U) THEN PRINT
  'ERROR: NUMBER OF COLUMNS OF U EXCEEDS NUMBER OF ROWS';
DO I = 1 TO NCOL(NX);
  X1 = J(NX[I], 1, I);
  IF I = 1 THEN X = X1;
  ELSE X = X//X1;
  END;
X = DESIGN(X);
NTOT = NROW(Y);
WOBS = NCOL(Y);
BOBS = NCOL(X);
```

```
WOBS1 = WOBS - 1;
**********FORM SIGMA MATRIX AND VECTOR OF MEANS**********;
BHAT = INV(X`*X)*X`*Y;
MUHAT = SHAPE(BHAT, WOBS#BOBS);
SIGMA = J(WOBS#BOBS, WOBS#BOBS, 0);
DF = NX - 1;
DO I = 1 TO BOBS;
  SIGB = (Y#X[,I] - X[,J]*BHAT[I,])`*(Y#X[,I] - X[,J]*BHAT[I,])/DF[I];
  F = I#WOBS - WOBS1;
  L = I#WOBS;
  SIGMA[F:L, F:L] = SIGB/NX[I];
END;
**********CALCULATE TEST STATISTIC, DF, AND P-VALUE**********;
R = C@U`;
T = (R*MUHAT)`*INV(R*SIGMA*R`)*(R*MUHAT);
A = 0;
IMAT = I(WOBS);
DO I = 1 TO BOBS;
  QMAT = J(BOBS#WOBS, BOBS#WOBS, 0);
  F = I#WOBS - WOBS1;
  L = I#WOBS;
  QMAT[F:L, F:L] = IMAT;
  PROD = (SIGMA*R`)*INV(R*SIGMA*R`)*R*QMAT;
  A = A + (TRACE(PROD*PROD) + TRACE(PROD)**2)/DF[I];
END;
A = A/2;
DF1 = NROW(R);
DF2 = DF1#(DF1 + 2)/(3#A);
CVAL = DF1 + 2#A - 6#A/(DF1 + 2);
RESULTS = J(4, 1, 0);
RESULTS[1] = T/CVAL;
RESULTS[2] = DF1;
RESULTS[3] = DF2;
RESULTS[4] = 1 - PROBF(RESULTS[1], DF1, DF2);
**************************PRINT RESULTS**************************;
PRINT 'WELCH-JAMES APPROXIMATE DF SOLUTION';
PRINT 'CONTRAST MATRIX:';
PRINT R[FORMAT = 4.1],;
MUHAT = MUHAT`;
PRINT 'MEAN VECTOR:';
PRINT MUHAT[FORMAT = 10.4],;
PRINT 'SIGMA MATRIX:';
PRINT SIGMA[FORMAT = 10.4],;
RESLAB = {"TEST STATISTIC" "NUMERATOR DF" "DENOMINATOR DF" "P-VALUE"};
PRINT 'SIGNIFICANCE TEST RESULTS:';
PRINT RESULTS[ROWNAME = RESLAB FORMAT = 10.4]/;
**************************END OF MODULE**************************;
FINISH;
```

At this point, the SAS/IML code needed to run the program for a particular research design is input.

Received February 10, 1994
Revision received October 18, 1994
Accepted October 19, 1994 ■