Journal of Modern Applied Statistical Methods

Volume 3 | Issue 2 Article 7

11-1-2004

Multivariate Contrasts For Repeated Measures Designs Under Assumption Violations

Lisa M. Lix
University of Manitoba, lisa.lix@usask.ca

Aynslie M. Hinds
University of Manitoba, umhinds0@cc.umanitoba.ca

 Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation
Lix, Lisa M. and Hinds, Aynslie M. (2004) "Multivariate Contrasts For Repeated Measures Designs Under Assumption Violations," *Journal of Modern Applied Statistical Methods*: Vol. 3 : Iss. 2 , Article 7.
DOI: 10.22237/jmasm/1099267620

Journal of Modern Applied Statistical Methods
November, 2004, Vol. 3, No. 2, 333-344
Copyright © 2004 JMASM, Inc.
1538 – 9472/04/$95.00

Multivariate Contrasts For Repeated Measures Designs Under Assumption Violations

Lisa M. Lix
Department of Community Health Sciences
University of Manitoba

Aynslie M. Hinds
Department of Community Health Sciences
University of Manitoba

Conventional and approximate degrees of freedom procedures for testing multivariate interaction contrasts in groups by trials repeated measures designs were compared under assumption violation conditions. Procedures were based on either least-squares or robust estimators. Power generally favored test procedures based on robust estimators for non-normal distributions, but was influenced by the degree of departure from non-normality, definition of power, and magnitude of the multivariate effect size.

Key words: à priori contrasts, robust estimators, covariance heterogeneity

Introduction

In a doubly multivariate repeated measures (RM) design, subjects provide data at K successive points in time or for each of K experimental conditions on p dependent variables. For example, measures of physical, social, psychological, and spiritual quality of life may be collected at multiple occasions during a course of treatment or therapy. A grouping factor (i.e., experimental vs. control group) is often included, resulting in a multivariate design in which both within-subjects main and interaction effects can be tested.

One approach for analyzing multivariate RM design is to follow statistically significant multivariate omnibus tests with multiple post hoc contrasts. Typically researchers will examine either strongly restricted contrasts, which are defined on the between- and/or within-subjects factor levels for a single dependent variable, or moderately restricted contrasts, which are defined on between-subjects and/or within-subjects factor levels for two or more dependent variables (Elliot & Barcikowski, 1994). A multivariate simultaneous test procedure (STP) will control the familywise error rate (FWR), the probability of making at least one erroneous decision regarding the null hypothesis for the contrasts, to the nominal level of significance, α (Bird & Hadzi-Pavlovic, 1983; Elliot & Barcikowski, 1994). The FWRs for both types of contrasts tend to be well below the nominal level of significance, α. In addition, these contrasts may have low power to detect effects in multivariate designs.

An alternate approach is to bypass the omnibus test in favor of à priori multivariate contrasts which test focused hypotheses on the between-subjects or within-subjects factor levels for a linear combination of the dependent variables (Huberty, 1994; Huberty, Chou, & Benitez, 1994; Keselman et al., 1998; Krishnaiah & Reising, 1985). These multivariate contrasts enable researchers to draw conclusions about the localized source of an effect while taking account of the correlation across repeated measurements and dependent variables.

One issue in conducting these a priori contrasts in multivariate designs is controlling the FWR. Timm (2002) has recommended the

Lisa Lix is Assistant Professor in the Department of Community Health Sciences. Her research interests include multivariate methods and longitudinal data/repeated measurements. E-mail: lisa_lix@cpe.umanitoba.ca. Aynslie Hinds is a graduate student in the Department of Community Health Sciences at the University of Manitoba Email: umhinds0@cc.umanitoba.ca.

use of STPs based on the Bonferroni inequality or the studentized maximum modulus. If the development of confidence intervals for these multivariate contrasts is not of primary concern, a stepwise procedure may also be considered (Keselman, Lix, & Kowalchuk, 1998; Tamhane & Dunnett, 1999).

A second issue is the choice of a test statistic and its associated derivational assumptions. In multivariate between-subjects designs, it is known that conventional procedures for testing post hoc contrasts are sensitive to violations of the assumptions of normality and covariance heterogeneity, which underlie the usual multivariate analysis of variance (MANOVA) tests (Bird & Hadzi-Pavlovic, 1983; Sheehan-Holt, 1998). In multivariate RM designs, the two conventional approaches for testing effects are the multivariate mixed model (MMM) and doubly multivariate model (DMM) approaches (Thomas, 1983; Boik, 1988, 1991).

The MMM rests on the stringent assumption of multivariate sphericity (M-sphericity). M-sphericity is the assumption that all pairwise differences of the repeated measurements exhibit a common variance for all dependent variables. In addition, both the MMM and DMM approaches rest on the assumptions of homogeneity of the covariances across between-subjects factor levels and multivariate normality. Because M-sphericity is not likely to be satisfied in practice, the DMM approach has been recommended over the MMM approach. However DMM tests are sensitive to violations of the assumptions of covariance homogeneity and multivariate normality.

The purpose of this article is to compare the conventional DMM procedure to procedures that employ approximate degrees of freedom (ADF) multivariate test statistics that do not rest on the assumption of covariance homogeneity for testing multivariate contrasts in repeated measures designs. Recent research (Lix, Algina, & Keselman, 2003; Lix, Keselman, & Hinds, in press) has derived multivariate ADF tests using robust estimators instead of the usual least-squares estimators which are known to be sensitive to departures from multivariate normality. Thus, it should be possible to obtain a test for multivariate contrasts in RM designs that is robust to both covariance heterogeneity and multivariate non-normality in multivariate RM designs, while controlling the FWR to α.

Definition of Test Statistics

Let $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Y}$= $[Y_{ijkl}]$, and Y_{ijkl} is the score for the ith individual (i = 1 ,…, n_j; $\sum_{j=1}^{J} n_j = N$) in the jth group (j = 1 ,…, J), on the kth (k = 1 ,…, K) repeated measurement and lth dependent variable (l = 1 ,…, p). Then $\mathbf{X}$ is an N x J design matrix with rank($\mathbf{X}$) = J, $\boldsymbol{\beta}$ is a J x L ($L = K$ x p) matrix of nonrandom parameters (i.e., population means), and $\boldsymbol{\varepsilon}$ is an N x L matrix of random error components. Each row of $\mathbf{Y}$ contains the L-dimensional response vector where the first K columns correspond to the repeated measurements obtained on the first dependent variable, the next K columns correspond to the repeated measurements obtained for the second dependent variable, and so on.

The null hypothesis for a multivariate contrast is

$$H_0 : \boldsymbol{\psi} = \mathbf{c\beta m} = \mathbf{0}, \qquad (1)$$

where $\mathbf{c}$ is a vector that contains the contrast coefficients for the between-subjects effect and $\mathbf{m} = (\mathbf{l} \otimes \mathbf{u}')$, where $\mathbf{l}$ = $\mathbf{I}_p$, the p x p identity matrix, $\otimes$ is the Kronecker product symbol, and $\mathbf{u}$ defines the contrast coefficients for the within-subjects effects. The best linear unbiased estimator for $\boldsymbol{\psi}$, which can be obtained by the least-squares method, is $\hat{\boldsymbol{\psi}} = \mathbf{c}\hat{\boldsymbol{\beta}}\mathbf{m}$, where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

Multivariate interaction contrasts are considered in this manuscript. There are a number of different types of interaction contrasts that may be defined for RM designs (Boik, 1993; Lix & Keselman, 1995). Tetrad contrasts, which are the simplest to define, test for differences between pairs of levels of two factors. For example, in a multivariate RM designs with J = 3 and K = 4, a tetrad contrast involving the first two levels of the between-subjects factor and the first and third levels of the within-subjects factor would require contrast

vectors of **c** = [1 -1 0] and **u**'=[1 0 -1 0].

Under a DMM approach, H_0 is tested with one of several well-known multivariate tests that are functions of the eigenvalues of $\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}$, where

$$\mathbf{H} = \hat{\boldsymbol{\psi}}\left[\mathbf{c}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{c}\right]^{-1}\hat{\boldsymbol{\psi}}', \tag{2}$$

and

$$\mathbf{E} = \mathbf{m}'\mathbf{Y}'\left[\mathbf{I}_N - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathbf{Y}\mathbf{m}, \tag{3}$$

where $\mathbf{I}_N$ is an identity matrix of dimension N. The tests are the Lawley-Hotelling trace, Pillai-Bartlett trace, Roy's largest root, and Wilk's lambda (Timm, 2002). If the multivariate contrast is a single-degree of freedom contrast on the p dependent variables, then all of these procedures will reduce to Hotelling's (1931) T^2.

When covariances are heterogeneous, Keselman and Lix (1997) demonstrated that DMM tests will produce inflated Type I error rates for omnibus tests of multivariate within-subjects effects, particularly when group sizes are unequal. Keselman and Lix (1997) showed than an ADF multivariate Welch-James (WJ) procedure due to Johansen (1980) can be used to test multivariate within-subjects main and interaction effects under covariance heterogeneity provided that sample sizes are sufficiently large. Moreover, Vallejo, Fidalgo, and Fernandez (2001) and Lix, Algina, and Keselman (2003) also demonstrated that a multivariate extension of the Brown and Forsythe (BF; Brown & Forsythe, 1974) procedure could be used to test within-subjects omnibus effects. The advantage of one procedure over the other depends on the omnibus effect of interest, total sample size, and the degree of covariance heterogeneity in the data.

Let $\mathbf{S}_j$ represent the sample covariance matrix for the jth group,

$$\mathbf{W}_j = \frac{\mathbf{m}'\mathbf{S}_j\mathbf{m}}{n_j}, \tag{4}$$

and $\mathbf{W} = \sum_{j=1}^{J}\mathbf{W}_j$. The WJ test statistic is

$$T_{\mathrm{WJ}} = \hat{\boldsymbol{\psi}}\mathbf{W}^{-1}\hat{\boldsymbol{\psi}}'. \tag{5}$$

The statistic T_{WJ}/C, where $C = p + 2A - 6A(p + 2)$, is distributed as $F_\alpha[\nu_{\mathrm{WJ1}},\nu_{\mathrm{WJ2}}]$, the $(1 - \alpha)$ percentile of the F distribution with $\nu_{\mathrm{WJ1}} = p$, $\nu_{\mathrm{WJ2}} = p(p + 2)/3A$, and

$$A = \sum_{j=1}^{J}\left[tr\left(\mathbf{I}_K - \mathbf{W}^{-1}\mathbf{W}_j\right)^2\right] + \left[tr\left(\mathbf{I}_K - \mathbf{W}^{-1}\mathbf{W}_j\right)\right]^2 / 2(n_j - 1). \tag{6}$$

For the BF procedure, define

$$\mathbf{W}_j^* = \frac{\left(1 - n_j\right)}{N}\mathbf{m}'\mathbf{S}_j\mathbf{m}, \tag{7}$$

and $\mathbf{W}^* = \sum_{j=1}^{J}\mathbf{W}_j^*$. The test statistic is

$$T_{BF} = \left(\frac{\nu_{\mathrm{h}}^*}{\nu_{\mathrm{e}}^*}\right)\hat{\boldsymbol{\psi}}\mathbf{W}^{*-1}\hat{\boldsymbol{\psi}}', \tag{8}$$

which is distributed as $F_\alpha[\nu_{\mathrm{BF1}},\nu_{\mathrm{BF2}}]$. The computations for ν_{BF1} and ν_{BF2} are lengthy, and the reader is referred to Vallejo et al. (2001) and Lix, Algina, and Keselman (2003) for the appropriate formulas.

Lix, Algina, and Keselman (2003) examined the WJ and BF procedures when least-squares estimators are replaced with robust estimators based on trimmed means. To define these procedures, let $Y_{(1)jkl} \le Y_{(2)jkl} \le \cdots \le Y_{(n_j)jkl}$, represent the ordered observations associated with the jth level of the between-subjects factor, the kth level of the repeated measures factor and the lth dependent variable. Let $g_j = [\gamma n_j]$, where γ represents the proportion of observations that are to be trimmed in each tail of the distribution and $[x]$ is the greatest integer $\le x$. The effective sample size for the jth group is $h_j = n_j - 2g_j$. The trimmed mean is estimated by

$$\hat{\mu}_{tjkl} = \frac{1}{h_j}\sum_{i=g_j+1}^{n_j-g_j} Y_{(i)jkl}. \tag{9}$$

Wilcox (1995a,b) has recommended that 20 percent trimming be adopted. It should be noted

that this is a univariate perspective on trimming, in which the most extreme scores for each column of **Y** are trimmed independently of the extreme scores in each of the other columns.

In order to obtain the sample Winsorized covariances, the sample Winsorized mean must first be computed and it is obtained by replacing the g_j smallest values with the γ percentile score, and the g_j largest values with the (1 - γ) percentile score

$$\hat{\mu}_{\mathrm{wj}kl} = \frac{1}{n_j}\sum_{i=1}^{n_j} Z_{ijkl}\,, \tag{10}$$

where

$$\begin{aligned} Z_{ijkl} &= Y_{(g_j+1)jkl} \text{ if } Y_{ijkl} \le Y_{(g_j+1)jkl} \\ &= Y_{ijkl} \text{ if } Y_{(g_j+1)jkl} < Y_{ijkl} < Y_{(n_j-g_j)jkl} \\ &= Y_{(n_j-g_j)jkl} \text{ if } Y_{ijkl} \ge Y_{(n_j-g_j)jkl}. \end{aligned}$$

The sample Winsorized covariance is required to obtain a theoretically valid estimate of the standard error of a trimmed mean. The covariance matrix of the Winsorized sample, $\mathbf{S}_{\mathrm{w}j} = \left[\hat{\sigma}_{\mathrm{w}jqq'}\right]$, is

$$\hat{\sigma}_{\mathrm{w}qq'} = \frac{\sum_{i=1}^{n_j}\left(Z_{ijqq'} - \hat{\mu}_{\mathrm{w}jqq'}\right)\left(Z_{ijqq'} - \hat{\mu}_{wjqq'}\right)}{\left(n_j - 1\right)}, \tag{11}$$

for $q, q' = 1, \ldots, L$.

To control the FWR for multiple tests, either a STP or a stepwise procedure may be adopted. For univariate RM designs under assumption violations, Lix and Keselman (1995) showed that the latter are more powerful, and recommended the use of either a step-up or step-down procedure based on the Bonferroni inequality, such as Hochberg's (1988) test.

Under Hochberg's procedure, one begins by rank ordering the *p*-values corresponding to the statistics used for testing the hypotheses $H_{(1)}, \ldots, H_{(B)}$, so that $p_{(1)} \le p_{(2)} \le \ldots \le p_{(B)}$ represent the ordered p-values. The decision rule is to reject $H_{(m')}$ ($m' \le m$; $m = B, \ldots, 1$) if $p_{(m)} \le \alpha/(B - m + 1)$. Testing begins with the hypothesis corresponding to the largest p-value, $p_{(B)}$. If $p_{(B)} \le \alpha$, all B hypotheses are rejected; if not, $H_{(B)}$ is retained and testing moves to $H_{(B-1)}$. If $p_{(B-1)} \le \alpha/2$, $H_{(C-1)}$ is rejected, as are all remaining hypotheses; if not $H_{(B-1)}$ is also retained, and $p_{(B-2)}$ is compared to $\alpha/3$, and so on. This continues, if all previous hypotheses have been retained, until $p_{(1)}$ is compared to α/B.

Methodology

A Monte Carlo study was used to evaluate the Type I error and power of the DMM, WJ and BF procedures for multivariate interaction contrasts. These three tests were investigated for a multivariate repeated measures design containing a single between-subjects factor with $J = 3$ levels and a single within-subjects factor with $K = 4$ levels.

The following variables were manipulated in the study. These were: (a) number of dependent variables, (b) total sample size, (c) equality/inequality of the group sizes, (d) the coefficient of variation of the group sizes for unbalanced designs, (e) degree of equality/inequality of the group covariance matrices, (f) nature of the pairing of group sizes and group covariance matrices, (g) multivariate normality/nonnormality, and (h) the non-null hypothesis for power comparisons. The degree of correlation between the dependent variables was set at $\rho = .80$. Keselman and Lix (1997) included both small and large p and ρ in their study; the former increased the total sample size required to obtain a robust solution for the WJ procedure, while the latter variable had little influence on the Type I error performance of the WJ procedure, which is consistent with previous research (Keselman & Lix, 1997). The pooled covariance of the repeated measurements had a non-spherical structure, with a value for ε, the index of non-sphericity, of $\varepsilon = .57$. The pooled covariance matrix had an average variance of 1.0 and average covariance of 0.5.

The procedures were investigated for $p = 2$ and 4 dependent variables for total sample sizes ranging from 60 to 120. The WJ test is likely to perform less optimally for small to

moderate sample sizes, particularly for non-normal distributions (Keselman et al., 2000). Both balanced and unbalanced designs were included in the study. For unbalanced designs, the sample size conditions were selected based on previous research (Keselman & Lix, 1997; Vallejo et al., 2001; Lix, Algina, & Keselman, 2003). Table 1 contains the values of the total sample sizes that were examined, along with the values of the coefficient of variation of the group sizes, Δn_j, where

$$\Delta n_j = \frac{\sqrt{\sum_{j=1}^{J}\left(n_j - \bar{n}\right)^2 / J}}{\bar{n}} . \qquad (12)$$

Table 1. Group Sizes (n_js) and Coefficient of Variation of Group Sizes (Δn_j) for Balanced and Unbalanced Designs.

N	n_j	Δn_j
60	20, 20, 20	0
	18, 20, 22	.08
90	30, 30, 30	0
	24, 30, 36	.16
	18, 30, 42	.33
120	40, 40, 40	0
	30, 40, 50	.20
	24, 40, 56	.33
	18, 40, 62	.45

This coefficient ranged in value from .08 to .45 when group sizes were unequal.

The procedures were investigated when the group covariance matrices were equal and unequal. For the latter case, the elements of the group covariance matrices were in a 1:3:5 ratio. These conditions are consistent with those selected by Keselman and Lix (1997) and Vallejo et al. (2001).

Both positive and negative pairings of group sizes and covariance matrices were investigated. A positive pairing refers to when the largest n_j is associated with the covariance matrix containing the largest element values; a negative pairing refers to the case in which the largest n_j is associated with the covariance matrix with the smallest element values.

Type I error and power rates were obtained when the data were both multivariate normal and non-normal in form. With respect to the former condition, pseudorandom observation vectors $\mathbf{Y}_{ij}$ from a multivariate normal distribution with mean vector $\boldsymbol{\beta}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$ were obtained using the SAS generator RANNOR (SAS Institute, 1999b). To obtain each $\mathbf{Y}_{ij}$, a row vector of L deviates in which each element has a standard normal distribution (i.e., $\mathbf{Z}_{ij}$), was transformed to a vector of multivariate observations via a triangular (Cholesky) decomposition, $\mathbf{Y}_{ij} = \boldsymbol{\beta}_j + \mathbf{L}\mathbf{Z}_{ij}^{\mathrm{T}}$, where $\mathbf{L}$ is a upper triangular matrix satisfying the equality $\mathbf{L}^{\mathrm{T}}\mathbf{L} = \boldsymbol{\Sigma}_j$. In this study, $\boldsymbol{\Sigma}_j$ was of the form $\boldsymbol{\Sigma}_j = (\boldsymbol{\Omega}_j \otimes \boldsymbol{\rho}_p)$ where $\boldsymbol{\rho}_p$ represent the p-dimension correlation matrix for the dependent variables and $\boldsymbol{\Omega}_j$ represents the K-dimension covariance matrix associated with a particular dependent variable for the jth group.

Two non-normal distributions were investigated: skewed and long-tailed. The skewed distribution had the same skewness (γ_1) and kurtosis (γ_2) values as a lognormal distribution, in which $\gamma_1 = 6.18$ and $\gamma_2 = 110.93$. The long-tailed distribution had skewness and kurtosis values equivalent to those of a double-exponential distribution, with $\gamma_1 = 0$ and $\gamma_2 = 3$.

These distributions and their associated measures of skewness and kurtosis are representative of those encountered in educational and psychological research (Micceri, 1989). The data were generated by the method developed by Fleishman (1978) and extended to the multivariate situation by Vale and Maurelli (1983).

For each distribution, a vector of constants, $\mathbf{w} = [a\ b\ c\ d]^{\mathrm{T}}$ was obtained using Fleishman's method, to provide the desired degree of multivariate skewness and kurtosis. An intermediate covariance matrix (i.e., $\boldsymbol{\lambda}_j$) was computed so that $\mathbf{Y}_{ij}$ would have the desired $\boldsymbol{\Sigma}_j$. Elements of this intermediate matrix were computed using Vale and Maurelli's (1983) Equation 11 (p. 467), which involves finding the roots of a third-degree polynomial; these roots were computed using the SAS/IML POLYROOT function (SAS Institute, 1999a). The vector of univariate standard normal

deviates was transformed to a vector of multivariate normal deviates via the Cholesky decomposition, $\mathbf{Z}(\boldsymbol{\lambda})_{ij} = \boldsymbol{\beta}_j + \mathbf{L}_\lambda \mathbf{Z}_{ij}^{\mathrm{T}}$, where $\mathbf{Z}(\boldsymbol{\lambda})_{ij}$ is the vector of transformed variates, and $\mathbf{L}_\lambda$ is an upper triangular matrix of dimension L satisfying the equality $\mathbf{L}_\lambda^{\mathrm{T}}\mathbf{L}_\lambda = \boldsymbol{\lambda}_j$.

Next, each element of $\mathbf{Y}_{ij}$ was obtained by computing the zero through third powers of the corresponding elements of $\mathbf{Z}(\boldsymbol{\lambda})_{ij}$, so that $\mathbf{Z}(\boldsymbol{\lambda})_{ijkl} = [1\ \mathbf{Z}(\boldsymbol{\lambda})_{ijkl}\ \mathbf{Z}(\boldsymbol{\lambda})_{ijkl}^2\ \mathbf{Z}(\boldsymbol{\lambda})_{ijkl}^3]$ which represents the vector of powers of the *kl*th components of $\mathbf{Z}(\boldsymbol{\lambda})_{ij}$. From this, $\mathbf{Y}_{ijkl} = \mathbf{Z}(\boldsymbol{\lambda})_{ijkl}\mathbf{w}$.

Three definitions of power were considered when non-null hypotheses were investigated. These were any-contrast power, that is, the power to detect at least one non-null hypothesis, all-contrast power, the power to detect all non-null contrasts, and average-per-contrast power, the average probability of detecting at least one non-null contrast. We examined the procedures when the effect size (f^2; Cohen, 1988) for the omnibus test of the within-subjects interaction was small and large for two patterns of non-null means.

For pattern 1, the first dependent variable had non-null means, while the second dependent variable had null means. For pattern 2, both dependent variables had the same non-null means. For patterns 1 and 2 respectively, the small effect size was equal to .16 and .08, respectively. The large effect size was 1.35 and .80 for patterns 1 and 2, respectively. The large effect size was selected to enable comparisons of all-contrast power across the investigated procedures; all-contrast power was zero for the small effect size.

The simulation program was written in the SAS/IML programming language (SAS Institute, 1999a). For the investigation of the FWR, the following factors were completely crossed: number of dependent variables (2), total sample size (3: small, moderate, large), relationship between group sizes and covariance matrices (4: equal group sizes/equal covariance matrices, equal group sizes/unequal covariance matrices, positive pairing of group sizes and covariance matrices, negative pairing of group sizes and covariance matrices), and population distribution (3: normal, double exponential, lognormal). The degree of sample size inequality was nested within total sample size.

For the investigation of power, the following factors were completely crossed for p = 2: total sample size, relationship between group sizes and covariances, population distribution, effect size (2: small, large), pattern of non-null means (2: non-null means on one dependent variable, non-null means on both dependent variables). For p = 2, five thousand replications of each condition were performed using a .05 significance level. For p = 4, because of the size of the matrices and the computations required, only three thousand replications were conducted. For each replication, the conventional DMM, WJ and BF tests were computed using least-squares and robust estimators.

Results

Type I Error

Table 2 contains the empirical percentages of FWR for the conventional (i.e., DMM), BF and WJ procedures for both least-squares and robust estimators for p = 2. Bold values are not contained within the bounds for Bradley's (1978) liberal criterion of robustness, which, for the five percent level of significance that was adopted, is 2.5 to 7.5 percentage points.

The data reveal that when the data were multivariate normal and least-squares estimators were adopted, the conventional test for multivariate contrasts could control the FWR when sample size was small or moderate for all conditions with the exception of negative pairings of group sizes and covariance matrices, and the positive pairing when Δn_j = .33. When sample size was large, the FWR was outside the bounds of Bradley's (1978) criterion for almost all of the positive and negative pairing conditions. When the data were normal, both the BF and WJ ADF procedures based on least-squares estimators controlled the rate of Type I errors across all of the investigated conditions.

Table 2. Empirical Percentages of Familywise Type I Error for Robust and Least Squares Estimators for Multivariate Interaction Contrasts, $p = 2$.

				Normal		Double Exponential		Lognormal	
N	Test	Pairing	Δn_j	LS	RE	LS	RE	LS	RE
60	DMM	$= n_j; = \boldsymbol{\Sigma}_j$	0	3.99	**2.34**	3.80	**2.22**	**1.66**	**2.40**
	BF			2.55	**0.89**	**2.29**	**0.78**	**0.44**	**0.78**
	WJ			3.85	**1.90**	3.35	**1.71**	**0.85**	**1.73**
	DMM	$= n_j; \neq \boldsymbol{\Sigma}_j$	0	7.11	4.72	6.72	4.25	2.76	4.56
	BF			3.24	**1.25**	2.85	**1.10**	**0.66**	**0.93**
	WJ			3.97	**1.93**	3.58	**1.85**	**1.03**	**1.66**
	DMM	+ pair	0.08	5.30	3.73	5.12	3.38	**2.01**	3.62
	BF			3.50	**1.45**	3.30	**1.22**	**0.65**	**1.25**
	WJ			3.78	**2.10**	3.44	**1.92**	**0.94**	**1.78**
	DMM	- pair	0.08	**9.13**	5.87	**7.88**	5.48	3.62	5.78
	BF			2.92	**1.08**	**2.43**	**0.92**	**0.53**	**0.94**
	WJ			3.92	**1.97**	3.17	**1.81**	**1.02**	**1.75**
90	DMM	$= n_j; = \boldsymbol{\Sigma}_j$	0	4.09	2.75	3.71	2.50	**1.93**	2.69
	BF			3.02	**1.46**	2.72	**1.31**	**0.75**	**1.34**
	WJ			3.88	**2.47**	3.50	**2.27**	**1.01**	**2.21**
	DMM	$= n_j; \neq \boldsymbol{\Sigma}_j$	0	6.95	5.20	6.44	4.93	3.07	5.59
	BF			3.75	**1.88**	3.40	**1.81**	**0.78**	**1.85**
	WJ			3.74	**2.37**	3.64	**2.22**	**1.00**	**2.37**
	DMM	+ pair	0.16	3.82	**3.18**	3.55	2.84	**1.59**	3.23
	BF			4.07	**2.35**	3.45	**2.03**	**1.08**	**1.90**
	WJ			3.76	2.63	3.27	**2.26**	**0.91**	**2.28**
	DMM	- pair	0.33	**2.30**	**1.79**	**2.37**	**1.71**	**1.04**	**1.79**
	BF			4.36	2.65	4.38	**2.48**	**1.57**	**2.36**
	WJ			3.70	**2.39**	3.58	**2.36**	**1.12**	**2.37**
	DMM	$= n_j; = \boldsymbol{\Sigma}_j$	0.16	**12.00**	**8.69**	**11.17**	**8.36**	5.86	**7.90**
	BF			3.54	**1.74**	3.15	**1.47**	**0.70**	**1.39**
	WJ			3.80	2.52	3.53	**2.19**	**1.06**	**2.14**
	DMM	$= n_j; \neq \boldsymbol{\Sigma}_j$	0.33	**18.84**	**14.73**	**18.26**	**14.72**	**10.85**	**14.35**
	BF			2.90	**1.38**	2.62	**1.15**	**0.60**	**1.01**
	WJ			3.80	**2.44**	3.59	**2.04**	**1.16**	**1.88**
120	DMM	$= n_j; = \boldsymbol{\Sigma}_j$	0	4.10	3.11	3.85	2.88	**2.15**	3.15
	BF			3.21	**1.89**	3.10	**1.76**	**1.00**	**1.74**
	WJ			4.01	2.78	3.66	**2.48**	**1.34**	2.68
	DMM	$= n_j; \neq \boldsymbol{\Sigma}_j$	0	7.06	5.65	6.34	5.12	3.30	5.58
	BF			4.12	**2.44**	3.47	**2.10**	**1.19**	**2.21**
	WJ			4.02	2.59	3.48	**2.30**	**1.30**	2.52
	DMM	+ pair	0.20	3.27	2.55	3.38	2.58	**1.72**	2.79
	BF			4.43	2.74	4.33	2.71	**1.63**	2.51
	WJ			4.09	2.62	3.71	2.60	**1.17**	2.73
	DMM	- pair	0.33	**2.38**	**1.93**	**2.24**	**1.90**	**1.09**	**2.21**
	BF			4.64	3.29	4.47	2.95	**1.84**	2.93
	WJ			3.98	2.72	3.63	2.64	**1.27**	2.73
	DMM	$= n_j; = \boldsymbol{\Sigma}_j$	0.45	**1.50**	**1.30**	**1.51**	**1.26**	**0.99**	**1.33**
	BF			4.78	3.22	4.62	3.05	**2.61**	3.26
	WJ			3.74	2.62	3.43	**2.48**	**1.35**	2.64
	DMM	$= n_j; \neq \boldsymbol{\Sigma}_j$	0.20	**13.43**	**10.85**	**13.12**	**10.85**	7.37	**11.00**
	BF			3.45	**1.95**	3.25	**1.71**	**0.88**	**1.47**
	WJ			3.90	**2.28**	3.36	**2.15**	**1.05**	**2.07**
	DMM	+ pair	0.33	**19.41**	**15.10**	**18.49**	**15.67**	**11.59**	**14.54**
	BF			3.54	**1.70**	2.90	**1.45**	**0.62**	**1.29**
	WJ			3.99	**2.44**	3.35	**2.13**	**1.18**	**2.23**
	DMM	- pair	0.45	**26.89**	**23.04**	**26.11**	**23.11**	**17.83**	**20.58**
	BF			3.07	**1.45**	**2.37**	**1.15**	**0.52**	**1.12**
	WJ			3.64	**2.29**	3.33	**2.00**	**1.26**	**2.18**

Note: + pair = positive paring of group sizes and covariance matrices; - pair=negative paring of group sizes and covariance matrices. Bold values are outside the range 2.5 - 7.5. LS = Least Squares estimators; RE=Robust estimators.

When the data were normal and robust estimators were adopted, the DMM test remained liberal for negative pairing conditions when sample size was moderate or large. The DMM, BF, and WJ procedures were frequently conservative, but this degree of conservatism decreased as the total sample size increased.

The results for symmetric and skewed distributions were substantially different. For the symmetric double exponential distribution, the FWR results for least-squares estimators was similar to those obtained for the normal distribution. That is, the DMM test was liberal for all negative pairing conditions and conservative for positive pairings when the degree of group size imbalance was large. The FWR for the ADF tests was well controlled. The same liberal tendencies of the DMM test were observed even when robust estimators were adopted, while the ADF tests were frequently conservative.

When the data were obtained from the skewed lognormal distribution, the error rates for the conventional and ADF procedures based on least-squares estimators were almost always conservative, except for negative pairings of group sizes and covariances when the DMM test could be liberal. When robust estimators were adopted, the FWRs for the DMM test could still be liberal. Those for the ADF tests tended to be less conservative than when least-squares estimators were adopted, and became even less so as total sample size increased.

The results for $p = 4$ (not reported) were similar to those provided in Table 2. However, the FWR for the conventional test were even more inflated than when $p = 2$. For example, when $N = 120$ and $\Delta n_j = .20$, the FWR was 18.14 and 12.32 percent for the double exponential distribution for least-squares and robust estimators, respectively.

Power

Table 3 contains the empirical percentages of any-contrast and average-per-contrast power for conventional and ADF procedures for the first mean pattern when the effect size was small. The data are averaged over all total sample size conditions. For the second mean pattern, any-contrast power attained its upper bound across most of the conditions; therefore these data are not reported. To interpret these results, we describe mean power differences of less than ten percentage points as small, between ten and 20 percent as moderate, and those of greater than 20 percent as substantial.

When the data followed a multivariate normal distribution, procedures based on least-squares estimators were more powerful than those based on robust estimators. The differences in any-contrast power were moderate to large. For average-per-contrast power they were small to moderate. For positive and negative pairing conditions, the differences in any-contrast power for the BF and WJ procedures were small; for positive pairings the BF test was slightly more powerful than the WJ test.

When the data had a multivariate heavy-tailed distribution, any-contrast power and average-per-contrast power rates for the procedures based on least-squares estimators were larger than those based on robust estimators. However, the differences were generally small. The exception was for any-contrast power for the BF and WJ procedures for negative pairings of group sizes and covariances, where the differences were moderate.

However, when the data had a multivariate skewed distribution, the procedures based on robust estimators were consistent more powerful than those based on least-squares estimators. This held true for both any-contrast and per-contrast power. The power differences were small to moderate. The WJ procedure was more powerful than the BF test with robust estimators across all of the investigated conditions.

Table 4 provides all-contrast and average-per-contrast power when the effect size was large for both the first and second mean patterns. Again, when the data followed a multivariate normal distribution, procedures based on least-squares estimators were always more powerful than those based on robust estimators. The differences in all-contrast power between the procedures based on least-squares estimators and those based on robust estimators were moderate to large for the first mean pattern.

Table 3. Empirical Percentages of Power for Robust and Least Squares Estimators for Multivariate Interaction Contrasts, $p = 2$; Small Effect Size, Mean Pattern 1.

		Normal				Double Exponential				Lognormal			
		ANCP		PCP		ANCP		PCP		ANCP		PCP	
Test	Pairing	LS	RE	LS	RE	LS	RE	LS	RE	LS	RE	LS	RE
DMM	$= n_j; = \boldsymbol{\Sigma}_j$	76.44	53.03	22.10	10.04	76.27	69.86	21.45	16.66	77.17	88.43	19.04	26.67
BF		72.17	44.28	20.15	7.81	71.66	61.12	19.31	13.44	68.60	81.57	14.71	21.39
WJ		74.92	48.72	20.96	8.77	74.95	65.66	20.45	14.82	79.24	85.06	19.55	24.07
DMM	$= n_j; \neq \boldsymbol{\Sigma}_j$	78.49	57.99	22.01	10.93	78.64	73.32	21.53	17.01	79.90	89.14	20.02	26.00
BF		70.29	42.59	17.80	6.96	70.16	58.57	17.22	11.62	68.57	77.42	14.51	17.89
WJ		66.53	37.89	19.42	6.79	66.54	54.54	18.93	12.36	73.59	85.44	18.93	21.34
DMM	+ pair	80.61	62.00	20.03	11.00	80.70	78.25	19.59	17.18	81.56	93.17	17.98	25.24
BF		81.19	61.93	22.00	11.76	81.40	76.55	21.44	17.67	80.66	90.55	17.94	24.62
WJ		75.95	53.25	24.51	10.74	76.19	70.47	23.90	18.27	80.06	87.69	22.29	29.40
DMM	- pair	83.19	67.15	28.72	15.72	82.86	79.49	27.97	23.14	83.27	93.18	25.40	35.05
BF		62.12	31.13	14.10	4.48	60.66	45.59	13.65	7.73	61.42	69.90	12.51	14.37
WJ		63.23	34.63	16.99	5.59	63.13	50.91	16.63	10.26	72.30	75.12	18.02	19.41

Note: ANCP = Any-contrast power; PCP = average-per-contrast power; LS = Least-squares estimators; RE = robust estimators.

The greatest difference was for the WJ procedure for both the positive and negative pairing conditions, where the difference in power was 25.5 and 27.2 percent for the positive and negative pairing conditions, respectively. The smallest difference was for the BF procedure. For the second mean pattern, the differences between least-squares and robust estimators were small to moderate when the data were normally distributed. Again, the greatest differences were for the WJ procedure. For per-contrast power, the differences between least-squares and robust estimators were small to moderate. The largest difference for both mean patterns was for the BF procedure (13.1 percent) when group sizes and covariance matrices were negatively paired.

For normally distributed data with least-squares estimators, the differences among the procedures varied considerably depending on the relationship between the group sizes and covariances. When the design was balanced and covariances were unequal, the WJ procedure was substantially more powerful than the BF procedure, and moderately more powerful than the DMM. The difference in power between the BF and WJ procedures was substantial for both the positive and negative pairing conditions for both mean patterns. This same pattern was evident when robust estimators were adopted.

For the double exponential distribution, the difference between procedures based on least-squares and robust estimators were small for both all-contrast and per-contrast power. The procedures based robust estimators were more power than those based on least-squares estimators for both types of power. Again, the differences between procedures based on least-squares and robust estimators were largest for the BF procedure.

Table 4. Empirical Percentages of Power for Robust and Least Squares Estimators for Multivariate Interaction Contrasts, $p = 2$; Large Effect Size.

		Normal				Double Exponential				Lognormal			
		ACP		PCP		ACP		PCP		ACP		PCP	
Test	Pairing	LS	RE	LS	RE	LS	RE	LS	RE	LS	RE	LS	RE
						Mean Pattern 1							
DMM	$= n_j; = \mathbf{\Sigma}_j$	27.18	8.39	83.72	75.02	25.93	19.93	83.48	81.58	17.54	37.66	81.06	26.67
BF		23.23	5.35	81.71	71.53	21.56	14.16	81.32	78.27	10.59	26.94	76.95	21.39
WJ		25.55	6.96	82.98	73.37	24.32	17.22	82.77	80.10	18.28	33.45	80.87	24.07
DMM	$= n_j; \neq \mathbf{\Sigma}_j$	27.34	7.43	83.20	73.92	25.81	19.73	82.90	80.96	18.00	38.88	80.77	26.00
BF		20.07	3.14	79.44	68.02	18.35	10.66	79.03	75.30	9.87	20.66	75.56	17.89
WJ		41.99	16.87	87.70	77.13	40.75	33.02	87.52	85.14	33.45	56.71	85.74	21.34
DMM	+ pair	15.49	2.68	79.19	71.21	13.99	10.46	78.78	77.78	8.23	23.46	76.28	25.24
BF		19.06	3.13	79.95	70.93	17.06	11.48	79.48	77.56	7.21	21.43	75.61	24.62
WJ		42.30	16.80	88.52	80.02	41.26	34.10	88.37	86.73	33.53	60.78	86.74	29.40
DMM	- pair	46.56	22.23	89.11	81.53	45.16	40.49	88.90	87.86	35.92	67.46	86.81	35.05
BF		23.12	2.75	80.27	67.21	21.17	9.62	79.81	75.12	14.64	24.70	76.58	14.37
WJ		49.24	22.06	89.22	78.33	47.90	42.27	88.97	87.55	38.12	69.81	86.65	19.41
						Mean Pattern 2							
DMM	$= n_j; = \mathbf{\Sigma}_j$	13.20	6.81	80.22	75.33	14.07	17.73	80.65	81.86	32.86	66.41	87.23	94.94
BF		9.13	3.71	78.17	72.05	9.73	11.32	78.49	78.66	23.09	54.93	83.72	92.52
WJ		11.95	5.62	79.47	73.83	12.94	15.13	79.97	80.50	34.36	61.33	87.50	93.86
DMM	$= n_j; \neq \mathbf{\Sigma}_j$	11.89	5.58	79.37	74.29	12.24	16.60	79.70	81.27	33.98	69.45	87.45	95.40
BF		5.49	1.67	75.42	68.62	5.85	7.14	75.70	75.69	21.19	49.44	82.76	90.99
WJ		26.95	14.33	85.05	77.84	28.09	31.37	85.50	85.91	54.42	80.62	92.28	97.34
DMM	+ pair	3.71	1.67	74.72	71.32	4.08	7.37	75.15	77.50	19.22	59.08	82.89	93.86
BF		3.97	1.42	75.42	70.88	4.33	6.82	75.75	77.16	17.30	55.87	82.28	93.00
WJ		25.74	13.79	85.59	80.07	27.12	31.13	86.04	86.94	54.14	85.22	92.52	98.09
DMM	- pair	31.83	18.80	86.62	81.63	32.80	37.91	86.88	88.06	54.95	87.20	92.32	98.22
BF		6.10	1.35	76.14	67.78	6.90	5.99	76.57	75.39	28.14	57.49	84.40	93.19
WJ		34.33	17.84	87.18	79.21	35.76	40.28	87.57	88.47	60.65	89.10	93.43	98.56

Note: ACP = All-contrast power; PCP = average-per-contrast power; LS = Least-squares estimators; RE = robust estimators.

When the data were obtained from a skewed distribution, the results also favor the procedures based on robust estimators. For all-contrast power, the power differences were moderate to substantial. Moreover, the WJ procedure demonstrated substantially greater power than the BF procedure across most of the investigated conditions. It was also more powerful than the DMM test when the design was balanced but covariances were unequal, and for positive pairings of group sizes and covariances.

Conclusion

The purpose of this article was to examine procedures for conducting multivariate a priori contrasts in RM designs. Conventional tests for multivariate within-subjects effects are sensitive to violations of the assumptions of covariance homogeneity and multivariate normality. Approximate degrees of freedom procedures are an appealing alternative because they are robust to heterogeneous covariance matrices. Furthermore, these tests can be extended to the

case of non-normal data by substituting least-squares estimators with robust estimators which are insensitive to the presence of skewed distributions and/or extreme observations.

Consistent with results for omnibus tests of the interaction (Keselman & Lix, 1997), the data show that error rates of conventional tests of multivariate interaction contrasts can become inflated when the group with the smallest number of observations exhibits the greatest degree of heterogeneity. These tests can also become conservative when there is a positive relationship between group sizes and covariances. The liberal and conservative tendencies do not disappear as sample size increases, and they become exacerbated as the dimension of the data increases.

Approximate degrees of freedom procedures based on least-squares estimators will perform well under violations of covariance homogeneity. These procedures will never be liberal under departures from multivariate normality. They may lose a moderate amount of power compared to procedures based on least-squares estimators. For the moderate degree of kurtosis that characterized the double exponential distribution, the differences in power between the tests based on robust estimators and those based on least-squared estimators were negligible, but did not always favor robust estimators. This power difference depended on the magnitude of the effect, the nature of the non-null means, and the definition of power that was adopted by the researcher. When the data were obtained from skewed distributions, the procedures based on robust estimators demonstrated clear power advantages in terms of detecting all contrasts of interest. Average-per-contrast power and any-contrast contrast power also favored robust estimators.

Previous research suggests that the Welch-James procedure should be selected over the Brown-Forsythe test when covariances and group sizes are negatively paired (Vallejo et al., 2001; Lix, Algina, & Keselman, 2003), this recommendation does not hold for all of the conditions investigated in this simulation study.

The choice of a procedure for testing within-subjects effects in multivariate repeated measures designs is complex, and depends on a number of factors. In this article, we advocate testing a set of hypotheses that enable the researcher to identify the localized source of multivariate interaction between a grouping factor and a repeated measures factor. If the data are in fact multivariate normal, then there is a modest gain in power to be obtained from adopting least-squares estimators. If the data are non-normal, there are power advantages by adopted a multivariate procedure that is robust to covariance heterogeneity and multivariate non-normality, particularly when the data are skewed. Which robust procedure to adopt is a function of the magnitude of the effect and the pattern of the non-null means. In closing, it should be noted that a SAS/IML (SAS Institute, 1999a) program to implement the Welch-James procedure with robust estimators for a variety of univariate and multivariate designs is available in Keselman, Wilcox, and Lix (2003).

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.

Boik, R. J. (1988). The mixed model for multivariate repeated measures: Validity conditions and an approximate test. *Psychometrika*, *53*, 469-486.

Boik, R. J. (1991). Scheffe's mixed model for multivariate repeated measures: A relative efficiency evaluation. *Communication 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.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: Erlbaum.

Elliott, R. S., & Barcikowski, R. S. (1994). Simultaneous test procedures in exploratory multivariate analysis of variance. American Educational Research Association Conference (1994, New Orleans, Louisiana).

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521-532.

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.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*, 800-802.

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

Huberty, C. J. (1994). *Applied discriminant analysis*. New York: John Wiley.

Huberty, C. J., Chou, T. F., & Benitez, E. B. (1994). The study of multivariate group contrasts. *Advances in Social Science Methodology*, *3,* 123-138.

Keselman, H. J., & Lix, L. M. (1997). Analyzing multivariate repeated measures designs when covariance matrices are heterogeneous. *British Journal of Mathematical and Statistical Psychology*, *50*, 319-338.

Keselman, H. J., Lix, L. M., & Kowalchuk, R. K. (1998). Multiple comparison procedures for trimmed means. *Psychological Methods*, *3*, 123-141.

Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L. M., & Wilcox, R. R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. *British Journal of Mathematical and Statistical Psychology*, *53*, 175-191.

Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophysiology*, *40*, 586-596.

Krishnaiah, P. R., & Reising, J. M. (1985). Multivariate multiple comparisons. In D. L. Banks, C. B. Read, & S. Kotz (Eds.), *Encyclopedia of Statistical Sciences, 6,* 88-95. New York: Wiley & Sons.

Lix, L. M., Algina, J., & Keselman, H. J. (2003). Analyzing multivariate repeated measures designs: A comparison of two approximate degrees of freedom procedures. *Multivariate Behavioral Research*, *38*, 403-431.

Lix, L. M., & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. *Psychological Bulletin*, *117*, 547-560.

Lix, L. M., Keselman, H. J., & Hinds, A. (in press). A comparison of procedures for the multivariate Behrens-Fisher problem. *Computer Methods and Programs in Biomedicine*.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 154-166.

SAS Institute Inc. (1999a). *SAS/IML software, Usage and reference, version 8*. Cary, NC: Author.

SAS Institute Inc. (1999b). *SAS language reference: Dictionary, version 8*. Cary, NC: Author.

Sheehan-Holt, J. K. (1998). MANOVA simultaneous test procedures: The power and robustness of restricted multivariate contrasts. *Educational and Psychological Measurement*, *58*, 861-881.

Tamhane, A. C., & Dunnett, C. W. (1999). Stepwise multiple test procedures with biometric applications. *Journal of Statistical Planning and Inference*, *82*, 55-68.

Thomas, D. R. (1983). Univariate repeated measures techniques applied to multivariate data. *Psychometrika*, *48*, 451-464.

Timm, N. H. (2002). *Applied multivariate analysis*. New York: Springer.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*, 451-464.

Vallejo, G., Fidalgo, A., & Fernandez, P. (2001). Effects of covariance heterogeneity on three procedures for analyzing multivariate repeated measures designs. *Multivariate Behavioral Research*, *36*, 1-27.

Wilcox, R. R. (1995a). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research*, *65*, 51-77.

Wilcox, R. R. (1995b). Simulation results on solutions to the multivariate Behrens-Fisher problem via trimmed means. *The Statistician*, *44*, 213-225.