Psicothema
2000. Vol. 12, nº 4, pp. 701-711

ISSN 0214 - 9915 CODEN PSOTEG

An examination of the robustness of the modified Brown-Forsythe and the Welch-James tests in the multivariate Split-Plot designs

Guillermo Vallejo and J. Ramón Escudero
University of Oviedo

The aim of this paper is to evaluate the robustness of the Welch-James multivariate solution given by Johansen (1980), and the improved multivariate Brown-Forsythe (1974) procedure when covariance matrices are heterogeneous. The results indicate that when design is unbalanced and the data are multivariate normally distributed, both approaches show a good control of error rates for the within-subjects main effect. When normality and homogeneity assumptions are jointly violated, none of the procedures was able to control the error rates in all of the investigated conditions. With regard to the test of the interaction effect, our results indicate that the modified Brown-Forsythe procedure can effectively control the rate of Type I errors when dispersion matrices are heterogeneous, and also when the data are sampled from a skewed distribution. This finding held even when the degree of heterogeneity of the covariance matrices was varied across the design. The Welch-James test is not a adequate solution, since the sample sizes required to achieve robustness could be unreasonably large, particularly when the multivariate normality assumption is violated.

Un examen de la robustez de las pruebas Welch-James y Brown-Forsythe modificada en diseños multivariados split-plot. Mediante el presente trabajo se pretende evaluar la robustez de la solución multivariada Welch-James dada por Johansen (1980) y la versión mejorada del enfoque multivariado de Brown y Forsythe (1974) cuando las matrices de dispersión son heterogéneas. Los resultados indican que cuando el diseño es desequilibrado y los datos son extraídos desde una distribución normal ambos enfoques controlan adecuadamente las tasas de error asociadas con el efecto principal de las ocasiones de medida. Sin embargo, cuando se incumplen los supuestos de normalidad y homogeneidad, ningún procedimiento es capaz de proporcionar un control estricto de las tasas de error. Por lo que respecta a la interacción, los resultados ponen de relieve que el procedimiento modificado de Brown-Forsythe ejerce un control muy satisfactorio de las tasas de error cuando los datos se obtienen desde distribuciones sesgadas. Este resultado también se mantiene cuando se el grado de heterogeneidad de las matrices de covarianza se varia a lo largo del diseño. Bajo esta condición el procedimiento de Welch-James no constituye una solución adecuada, dado que los tamaños de muestra requeridos para lograr la robustez pueden llegar a ser exagerados, sobre manera, cuando los datos carecen de normalidad.

The univariate repeated measures design containing a single between-subjects (groups) factor *A* with $j=1,\ldots,p$ levels and n_j observations at each *j* and a single within-subjects (occasions) factor *B* with $k=1,\ldots,q$ levels is very frequent in almost all scientific fields (Shoukri & Pause, 1999). Although the nature of these designs is typically multivariate, the effects of design (occasions main effect and groups x occasions interaction) can be tested by using univariate or multivariate approach. The validity of these procedures rests on the nature of the assumptions that the researcher is willing to make about the data. When the assumptions of multivariate normality, homogeneity of the covariance matrices, and multisample sphericity are satisfied, such designs are analyzed by Scheffé's (1956) univariate mixed model. When the multisample sphericity assumption is not satisfied either an adjusted degrees of freedom univariate test or multivariate model perspective may be used. Under a multivariate model, no restrictions are placed on the structure of the covariance matrix. However, the number of experimental observations must be greater or equal to the of repeated measurements and, as the univariate model, the assumptions of dispersion matrix equality and normality must be satisfied.

If sphericity assumption is met, the conventional univariate procedure is more powerful than the multivariate approach (Davidson, 1972). However, if sphericity appears untenable no clear-cut rule emerged for choosing between the adjusted degrees of freedom univariate tests and their multivariate counterparts (Mendoza, Toothaker, & Nicewander, 1974). When covariance matrices are unequal and the design is balanced (equal group sizes), Keselman and Keselman (1990) have shown that both procedures are generally robust to the violation of dispersion matrix equality. In this case, the choice between univariate or multivariate technique depends, especially, on differences in their statistical power. Ho-

Correspondencia: Guillermo Vallejo
Facultad de Psicología
Universidad de Oviedo
33003 Oviedo (Spain)
E-mail: gvallejo@sci.cpd.uniovi.es

wever, none of these approaches can provide robust tests of within-subjects main and interaction effects when matrices are heterogeneous and group sizes are unequal. In this last situation, Algina and Oshima (1995) suggested using the General Approximation or Improved General Approximation test due to Huynh (1978), Keselman, Carriere, and Lix (1993) suggested using the multivariate Welch-James (WJ) statistic given by Johansen (1980), whereas Jones (1993) defends the mixed model approach. That is, rather than presuming a certain type of structure, as is the case with the univariate or a multivariate test statistic, the advocates the mixed model approach modeling the covariance structure directly. Yet, recently, Keselman, Algina, Kowalchuk and Wolfinger (1999) have found that this new approach, as implemented in *SAS* (Release 6.11 of PROC MIXED, SAS Institute, 1996), has some problems in identifying the correct structure, and, is times, it is prone to depressed or inflated Type I error rates. For this reason they suggest apply the approach cautiously, proposing as alternative solution the multivariate *WJ* statistic.

In a multivariate repeated measures experiment, each subject gives a *r*-dimensional response on each of *q* occasions. In this case, if the *r* variables are statistically related or if the Type I error rate is to be controlled experimentwise, either a multivariate mixed model (MMM; the Scheffé's mixed model generalized for application to multivariate case) or doubly multivariate model (DMM) perspective may be used. Both analyses require (a) that the populations sampled have a multivariate normal distribution, and (b) that the dispersion matrices are the same for the populations sampled. Otherwise, the *MMM* analysis rests upon a further assumption, namely multivariate sphericity (M-sphericity). Simulation studies have shown that the unadjusted *MMM* test cannot be recommended except when *M*-sphericity is known to hold. One situation in which the adjusted *MMM* test is more powerful than the *DMM* test is when sample size is very small. If sample size is reasonably large, there appears to be little or no advantage in using adjusted *MMM* tests. When the sample contains adequate information to estimate the covariance matrix without requiring any particular structural form, the *DMM* test must be preferred since almost always provide greater statistical power (Boik, 1991; Vallejo & Menéndez, 1997; Vallejo, Fidalgo, & Fernández, 1998).

Vallejo, Fernández, Fidalgo, and Escudero (1999) evaluated the power and robustness for the *DMM* test and the ε-corrected *MMM* test suggested by Boik (1991) in the presence of heteroscedasticity of the variance-covariance matrices and when data were non-normal in form under null and non-null hypothesis. Their results revealed that these tests were extremely sensitive to departures from covariance homogeneity when the design was unbalanced (unequal group sizes) and the sample size was small. When the design was balanced, both adjusted *MMM* and *MDM* approaches exhibited a superior control of error rates. Data distribution had small effects on the Type I error rates and power for both procedures: the *DMM* test was slightly liberal when the model was additive and conservative when the model was non-additive; its effect for corrected *MMM* tests was insignificant. These results are consistent with the empirical literature (Keselman & Keselman, 1990; Keselman & Lix, 1997; Mendoza *et al.* 1974; Olson, 1974; Rogan, Keselman, & Mendoza, 1980).

Subsequently, Vallejo, Fidalgo and Fernández (in press) evaluated the robustness of the doubly multivariate model, Welch-James multivariate solution and the multivariate version of the modified Brown-Forsythe (BF, 1974) procedure proposed by Rubin (1983) and Mehrotra (1997), within the context of one-way analysis of variance. The performance of these procedures was investigated by testing within-blocks sources of variation in unbalanced multivariate split-plot designs containing unequal covariance matrices. Our findings indicate that the doubly multivariate model did not provide effective Type I error control, while the Welch-James procedure provided robust and powerful tests of the within-subjects main effect; however, this approach provided liberal tests of the interaction effect. The results also indicate that the modified Brown-Forsythe procedure provided robust tests of within-subjects main and interaction effects, especially when the design was balanced, or when group sizes and covariance matrices were positively paired.

Vallejo *et al*. (in press) did not consider the effects of multivariate non-normality on the operating characteristics of the examined procedures. Thus, additional research is necessary to determine if the findings obtained by Vallejo *et al.* (in press) can be generalized beyond the limited conditions they investigated. In particular, it is very important to examine the robustness of modified *BF* procedure when the degree of heterogeneity of the covariance matrices is varied across the designs and the data are not normally distributed. Accordingly, the main purpose of this study is to compare the Type I error rates of the *WJ* and modified *BF* statistics for testing within-subjects main and interaction effects in multivariate repeated measures designs, in the presence of heteroscedasticity variance-covariance matrices and multivariate non-normality. A second purpose of this study is to determine if the *BF* test offers a greater control of Type I error rates for the interaction than the *WJ* when the sample sizes are sufficiently large.

Definition of Test Statistics

The linear model for multivariate repeated measures can be written as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}, \tag{1}$$

where $\mathbf{Y}$ is the $N \times qr$ response matrix, $\mathbf{B}$ is the $p \times qr$ matrix of parameters, $\mathbf{X}$ is the $N \times p$ design matrix of full rank, and $\mathbf{U}$ is the $N \times qr$ matrix of random errors. If ε_i' denotes a vector of random errors associated with the *ith* subject, it is assumed that $\varepsilon_i' \sim N(\mathbf{0}, \Sigma_j)$ where Σ_j is the $qr \times qr$ matrix of dispersion corresponding to the *jth* level of the between-subjects factor. Jointly,

$$vec(\mathbf{U}) \sim N[\mathbf{0}, (\mathbf{I}_N \otimes \Sigma_j)] \tag{2}$$

where the symbol ⊗ represents the direct or Kronecker product of two matrices. The fact that Σ_j depends upon *i* means that the covariance matrices for the repeated measures vary across groups.

Multivariate Brown-Forsythe (BF) test

The general linear hypothesis for the *BF* procedure can be written as

$$H_0 : \mathbf{C'BA} = \mathbf{0} \tag{3}$$

where $\mathbf{C}'$ is a $v_h \times p$ matrix of rank n_h, $\mathbf{B}$ was defined before, and $\mathbf{A}$ is a $q \times u$ matrix of rank u. Coefficients for between-subjects contrasts are contained in $\mathbf{C}$ and coefficients for within-subjects contrasts for the r dependent variables are contained in $\mathbf{A}$.

The *BF* statistics for testing the hypothesis concerning to the within-subjects interaction, assuming (2), can be expressed in terms of the matrices **H** and **E***. These matrices take the following form,

$$H = (C'\hat{B}A)'[C'(X^{-1}X)C]^{-1}(C'\hat{B}A) \tag{4}$$

and

$$E^* = \left(\frac{v_e^*}{v_h^*}\right)\sum_{j=1}^{p} c_j A'\Sigma_j A \tag{5}$$

where $\mathbf{C'} = [\mathbf{I}_{p-1} \vdots -1]$ and $\mathbf{A} = \mathbf{I}_r \otimes \mathbf{F}$, with $\mathbf{F} = [\mathbf{I}_{q-1} \vdots -1]$ and $c_j = (1 - n_j \mid N)$. This form of $\mathbf{E}^*$ matrix ensures that the expected values of **H** and the expected value of $\sum_{j=1}^{p} c_j A'\Sigma_j A$ are equal if the null hypothesis is true, since mean vectors are being compared across groups.

Using results in Nel and van der Merwe (1986), the distribution of matrix $\sum_{j=1}^{p} c_j A'\Sigma_j A$ can be approximated as a sum of Wisharts distribution

$$\sum_{j=1}^{p}(c_j A'\Sigma_j A) \sim S\,W_{qr}\left(v_1^*,\ldots,v_p^*;\frac{c_1}{v_1^*}A'\Sigma_1 A,\ldots,\frac{c_p}{v_p^*}A'\Sigma_p A\right) \tag{6}$$

with degrees of freedom

$$v_e^* = \frac{tr\left[\sum_{j=1}^{p}(c_j A'\Sigma_j A)\right]^2 + \left[tr\sum_{j=1}^{p}(c_j A'\Sigma_j A)\right]^2}{\sum_{j=1}^{p}\frac{1}{n_j-1}\left\{tr(c_j A'\Sigma_j A)^2 + \left[tr(c_j A'\Sigma_j A)\right]^2\right\}} \tag{7}$$

The symbol *tr* denotes the trace of a matrix.

This hypothesis was tested using the F-test approximation to Wilk's L given by Rao (1951) as

$$F = \frac{1-\Lambda^{1/s^*}}{\Lambda^{1/s^*}}\left(\frac{\mu_2^*}{\mu_1^*}\right), \tag{8}$$

where $s^* = [(m^2\mu_h^{*2} - 4)/(m^2 + \mu_h^{*2} - 5)]^{1/2}$, $\mu_1^* = m\mu_h^*$, $\mu_2^* = \{[\mu_e^* - (m - \mu_h^* + 1)/2]s^* - (m\mu_h^* - 2)/2\}$, $\mu_h^* = \mu_h.\mu_h/(p-1)$, and $\Lambda = |\mathbf{E}^*|/|\mathbf{E}^* + \mathbf{H}|$, with *m* equal to the dimension of $\mathbf{E}^*$ and **H** and μ_h^* equal to

$$v_h^* = \frac{2\{tr(MP)\}^2}{tr(MP)^2}, \tag{9}$$

where $\mathbf{M}=\mathbf{R}(\mathbf{R'GR})^+\mathbf{R'}$ and $\mathbf{P}=\mathbf{diag}\,[(n_1/N)^{-1}\Sigma_1,\ldots,(n_p/N)^{-1}\Sigma_p]$, with $\mathbf{R}=\mathbf{C'}\otimes(\mathbf{I}_r\otimes\mathbf{F'})$, $\mathbf{G}=\{[N^{-1}(\mathbf{X'X})]^{-1}\otimes\Lambda\}$, $\Lambda=\text{diag}\,(1_1, 0_2,\ldots, 0_{qr})$, and $(.)^+$ is the Moore-Penrose inverse of (.). This hypothesis was rejected at nominal α level if $F > F_{(1-\alpha);\,\mu_1^*,\,\mu_2^*}$, where $F_{(1-\alpha);\,\mu_1^*,\,\mu_2^*}$ is the 100 $(1-\alpha)$ *th* percentile of the F-distribution with μ_1^* and μ_2^* degrees of freedom.

The above result was established assuming that the quadratic form **H** can be approximated as weighted sum of Wisharts distribution

$$H \sim \sum_{j=1}^{m}\lambda_j W_j(v_j, P\Phi), \tag{10}$$

where each Wishart distribution in the sum has one degree of freedom and $\lambda_1, \lambda_2,\ldots,\lambda_m$ are distinct nonzero eigenvalues of **MP** (or **PM**). Taking $H = \sum_{j=1}^{m}\lambda_j W_j(v_j, P)$ Khatri (1980) find that

$$\text{E}(\mathbf{H}) = \lambda_j\,\mu_j, \tag{11}$$

and

$$\text{V}(\mathbf{H}) = 2\lambda_j^2\,\mu_j\,. \tag{12}$$

Equating the first two moments of the distribution of **H** $[E(\mathbf{H})=tr(\mathbf{MP})$ and $V(\mathbf{H})=tr(\mathbf{MP})^2]$ to those of a central Wishart matrix and solving simultaneously the equations (11) and (12) we obtained the equation (9).

The statistics used to test the within-subjects main effect hypothesis also can be expressed in terms of the matrices **H** and $\tilde{\mathbf{E}}$ where

$$H = (C'\tilde{B}A)'[C'(X^{-1}X)C]^{-1}(C'\tilde{B}A), \tag{13}$$

and

$$\tilde{E} = \left(\frac{\mu_e^*}{N^{-1}p^2}\right)\sum_{j=1}^{p} n_j^{-1}A'\Sigma_j A. \tag{14}$$

In equation (13), **C**' is a *1 x p* vector of ones, **A** is as previously defined, and $\tilde{\mathbf{B}} = (\bar{n}/\tilde{n})^{1/2}\,\hat{\mathbf{B}}$, where the symbols $\bar{n}$ and $\tilde{n}$ designates the arithmetic mean and the harmonic mean of n_j's, respectively. Extending the results reported by Nel and van der Merwe (1986), the distribution of matrix $\sum_{j=1}^{p} n_j^{-1}A'\Sigma_j A$ can be approximated as a sum of Wisharts distribution

$$\sum_{j=1}^{p} n_j^{-1}A'\Sigma_j A \sim S\,W_{qr}\left(v_1^\bullet,\ldots,v_p^\bullet;\frac{1}{n_1 v_1^\bullet}A'\Sigma_1 A,\ldots,\frac{1}{n_p v_p^\bullet}A'\Sigma_p A\right), \tag{15}$$

with degrees of freedom

$$v_e^\bullet = \frac{tr\left[\sum_{j=1}^{p}(n_j^{-1}A'\Sigma_j A)\right]^2 + \left[tr\sum_{j=1}^{p}(n_j^{-1}A'\Sigma_j A)\right]^2}{\sum_{j=1}^{p}\frac{1}{n_j-1}\left\{tr(n_j^{-1}A'\Sigma_j A)^2 + \left[tr(n_j^{-1}A'\Sigma_j A)\right]^2\right\}} \tag{16}$$

This hypothesis was tested using the F-test approximation to Wilk's L given by Rao as

$$F = \frac{1-\Lambda^{1/s}}{\Lambda^{1/s}}\left(\frac{\mu_2^\bullet}{\mu_1}\right), \tag{17}$$

where $s= [(m^2\mu^2_h - 4)/(m^2 + \mu^2_h - 5)]^{1/2}$, $\mu_1= m\mu_h$, and $\mu^\bullet_2 = \{[\mu^\bullet_e - (m - \mu_h + 1) / 2]s - (m\mu_h - 2) / 2\}$.

The Welch-James (WJ) test

The multivariate *WJ* statistic for testing repeated measures main and interaction effect hypotheses developed by Keselman *et al.* (1993) according to Johansen (1980), can be used when the covariance homogeneity assumption is not satisfied. The approximate degrees of freedom multivariate *WJ* type statistic is

$$T_{WJ} = \mathbf{(R\bar{y})' (RPR')^{-1} (R\bar{y})}, \tag{18}$$

where $\mathbf{y}$ is a *pqr x 1* vector with elements obtained by stacking the mean of $\mathbf{y}_j$, $\mathbf{R = C' \otimes (I_r \otimes F')}$ is a contrast matrix whose order depends on the hypothesis tested, and $\mathbf{P}$ is a block diagonal matrix of dimension *pqr x pqr* with the sample covariance matrices weighted by n^{-1}_j in the main diagonal. This test statistic, divided by a constant, *c*, can be approximate by an F distribution with μ_1 (rank of the $\mathbf{R}$ contrast matrix) and $\mu_2= \mu_1(\mu_1 + 2) /(3A)$. The constant $c= \mu_1 +2A - 6A /(\mu_1 +2)$, with

$$A = \frac{1}{2}\sum_{j=1}^{p}\left[tr\left\{PR'(RPR')^{-1}RQ_j\right\}^2 + \left\{tr(PR'(RPR'^{-1})RQ_j)\right\}^2\right] / (n_j - 1), \tag{19}$$

where $\mathbf{Q}_j$ is a block diagonal matrix of dimension *pqr x pqr*, with the *jth* block equal to a *qr x qr* identity matrix and zeroes elsewhere.

Vallejo and Escudero (1998) showed that for testing $H_0 : \mathbf{R\mu = 0}$, the form of the $\mathbf{R}$ matrix depends on the tested effect. For the interaction, $\mathbf{R = C' \otimes (I_r \otimes F')}$ where $\mathbf{C}'$ is a *(p - 1) x p* coefficient matrix that determines the elements of $\mathbf{B}$ include in the null hypothesis, $\mathbf{F}$ is a *q x (q - 1)* coefficient matrix for testing hypothesis about the repeated measures variable, and $\mathbf{I}_r$ is an *r x r* identity matrix. Whereas for the within-subjects main effect (additive model and unweighted means), $\mathbf{R = c' \otimes (I_r \otimes F')}$ where $\mathbf{c}'$ is a *1 x p* vector of ones, $\mathbf{F}$ is a *q x (q - 1)* contrast matrix, and $\mathbf{I}_r$ is an *r x r* identity matrix.

For both effects the $H_0 : \mathbf{R\mu = 0}$, is rejected using a significance level of α if $T_{WJ} / c > F_{(1-\alpha);\, \mu_1,\, \mu_2}$, where $F_{(1-\alpha);\, \mu_1,\, \mu_2}$ is the 100 $(1-\alpha)$ *th* percentile of the F-distribution with μ_1 and μ_2 degrees of freedom.

Method

A Monte Carlo simulation study was undertaken to evaluate the robustness of the *BF* and *WJ* statistics for testing within-subjects main and interaction effects. The design investigated herein had one between-subjects factor (*p* = 3), one within-subjects factor (*q* = 4), and three dependent variables (*r* = 3). Five variables were manipulated. These were: (a) total sample size (N), (b) nature of the pairing of unequal covariance matrices and group sizes, (c) types of population covariance structures, (e) degree of heterogeneity of the covariance matrices, and (e) types of distributions.

Based on the previous research findings, the first variable, *N*, was selected such that the ratio of *N / r (q-1)* was ranged from 8 to 16. Thus, for *r (q-1) = 9*, *N* = 72, 108, and 144. Though, unfortunately, the last value is not very frequent in the educational and psychological researches according to the survey conducted by Kowalchuk, Lix, and Keselman (1996), for comparison purposes we have adopted.

The second variable manipulated in the current investigation was pairing condition. Null, positive and negative pairing of group sizes and covariance matrices were investigated. A null pairing refers to the case in which matrices are heterogeneous but the design is balanced, that is, the size of the element values at the covariance matrices were not related with the group sizes because all groups had an equal size. A positive pairing referred to the case in which the largest n_j was associated with the covariance matrix containing the largest element values; a negative pairing referred to the case in which the largest n_j was associated with the covariance matrix containing the smallest element values. For positive a negative pairings, a moderate and substantial degree of group size inequality was investigated. The moderately unbalanced group sizes had a coefficient of sample size variation (Δ) equal to .20, while the more disparate cases Δ = .40, where

$$\Delta = \frac{1}{\bar{n}}\left[\sum_{j=1}^{p}\left(n_j - \bar{n}\right)^2 / p\right]^{1/2}, \tag{20}$$

and $\bar{n}$ is the average group size. When the design is balanced Δ = 0, whereas when the design is unbalanced this coefficient increase in value as group sizes become more disparate. Finally, the ratio of the smallest group size (e.g., n_{min}) to *r (q – 1)* were set at 1.33 for N = 72, 2 for N = 108, and 2.67 for N = 144.

The third variable investigated was the pattern of covariance matrices. In this study, the forms of the dispersion matrices were $\Sigma_j= (\Psi_r \otimes \mathbf{V}_j)$ and $\Sigma_j= (\Psi_r \otimes \mathbf{W}_j)$, where Ψ_r represents the *r x r* correlation matrix for the dependent variables, and $\mathbf{V}_j$ and $\mathbf{W}_j$ describes the covariance among the repeated measures associated with a particular dependent variable. In the first condition the matrix $\mathbf{V}_j$ had compound symmetry (CS), whereas in the second condition the matrix $\mathbf{W}_j$ had serial correlation (AR). Though the *BF* and *WJ* procedures are multivariate statistics and therefore should not be dependent of the pattern of covariance matrices, Vallejo *et al*. (1999) found that the rate of Type I error for the *DMM* test does vary with the form of Σ. In particular, if the covariance matrix has a Kronecker structure.

The fourth variable included in this study was the degree of heterogeneity of the covariance matrices. Two levels of dispersion matrix inequality were varied: ($\Sigma_1 = 1/3\Sigma_2$ and $\Sigma_3 = 5/3\Sigma_2$) and ($\Sigma_1 = 1/5\Sigma_2$ and $\Sigma_3 = 9/5\Sigma_2$).

The last variable investigated was the type of distribution. Type I error rates were obtained when the data were both normal and non-normal in form. With respect to the former condition, the data were generated as follows:

1. For each level of the between-subjects factor, generate vectors of pseudo-random normal variates. The *GAUSS* generator *RNDN* (GAUSS Aptech Systems, 1997) was used to obtain all vectors of normal variates.

2. The corresponding multivariate observations were obtained by the method of Schauer and Stoller (1966), that is, $\mathbf{y'}_{ij} = \mathbf{Lz_{ij}} + \boldsymbol{\mu}_{\mathbf{ij}}$, where $\mathbf{L}$ is a Cholesky factor of Σ_j and $\mathbf{z}_{ij}$ is a vector of normal variates obtained from the Kinderman and Ramage (1976) algorithm.

The non-normal data for the current study were sampled from a chi-squared distribution with three degrees of freedom as follows:

1. For each level of the between-subjects factor, to obtain each $\mathbf{w}_{ij}$, a vector of variates having a χ^2 distribution with three degrees of freedom, three vectors of pseudorandom normal variates we-

re squared and summed. The *GAUSS* generator *RNDN* (GAUSS Aptech Systems, 1997) was used to generate all variates.

2. The χ^2 variates generated in the precedent step were standardized to have a mean zero and variance one using the population expected value and standard deviation. See Hasting and Peacock (1975) for further details on the generation of data from this distribution.

3. The corresponding multivariate observations were obtained using the same procedure as was used for the normal distribution.

This particular type of $\chi^2_{(3)}$ distribution with γ_1 (skewness)=1.63 and γ_2 (kurtosis)=4 was selected for three reasons. First, Micceri (1989) investigated many data sets from educational and psychological research and found striking departures from normality. Second, this population represent relatively extreme but realistic skew-leptokurtic distribution (see, Micceri, 1989, Wilcox, 1989). Third, this population has been used in a number of previous studies designed to investigate of the robustness of the *WJ* procedure (p.e., Keselman *et al.,* 1993; Algina and Keselman, 1997).

The simulation program was written in the *GAUSS* programming language. All factors were completely crossed with one another: three sample sizes (72, 108, and 144), five patterns of pairings (one null, two positive, and two negative), two types of covariance structures, two levels of dispersion matrix inequality, and two types of distributions. For each of the 3 x 5 x 2 x 2 x 2 = 120 cells of the design the number of replications was 10,000. Using Wilk's (1932) lambda, the *BF* and *WJ* statistics for testing hypothesis concerning main and interaction effects were performed using the 0.05 and 0.01 nominal significance level. A summary of conditions included in the study is presented in Table1.

Results

Estimated Type I error rates ($\hat{\alpha}$) are reported in Table 2, in Table 3, in Table 4, and Table 5. On these tables, $\hat{\alpha}$ outside the interval $\alpha/2 \le \hat{\alpha} \le 3/2\alpha$ are in bold. According to this criterion, in order for a test to be considered robust, its empirical rate of Type I error must be contained in the interval ($.025 \le \hat{\alpha} \le .075$) for the 5% level of significance, and in the interval ($.005 \le \hat{\alpha} \le .015$) for the 1% level of significance. Correspondingly, a test was considered to be non-robust if, for a particular condition, its Type I error was not contained in these intervals. Although to evaluate the adequacy of robustness in control of Type I errors, several standards have been used, Keselman and Lix (1997) used this criterion and thus for comparison purposes we have adopted it as well. Nonetheless, it should be noted that with other standards different interpretations of the results are possible.

Type I Error Rates for Tests of the Occasions Main Effect

*Normally Distributed Data.*Table 2 contains the empirical rates of Type I error for the main effect of the *BF* and *WJ* tests for each manipulated condition.

As seen from table 2, the *WJ* statistic was able to control the Type I error rates across all of the investigated conditions, even when the sample sizes are small. Similar results were obtained with the *BF* procedure, except for negative pairing condition, when N = 72 and Δ = .40. In this case, the procedure was always conservative. The other two manipulated conditions, that is, covariance ratios and pattern of covariance matrices had little effect on the results associated with both procedures.

Nonnormally Distributed Data. Table 3 contains the empirical rates of Type I error for the main effect when data were sampled from a chi-squared distribution with three degrees of freedom.

As seen from table 3, when the data are obtained from a skewed distribution increases Type I error rates for the *BF* and *WJ* tests, in particular, for α = .01. For the *BF* test 4 conditions resulted in Type I error rates below .005, and 16 rates above .015. Whereas, for the *WJ* test 23 conditions resulted in Type I error rates above .005, and 8 rates above .075.

In this case, contrary to what happened when data were sampled from a multivariate normal distribution, covariance ratios and pattern of covariance matrices had a superior effect on the robustness of both procedures; especially, the degree of heterogeneity of the covariance matrices.

Table 1
Summary of experimental conditions

						Normal Data				Nonnormal Data			
						1/3: 1:5/3		1/3: 1:5/3		1/5: 1:9/5		1/5: 1:9/5	
N	**n_1,**	**n_2,**	**n_3**	**Pairing**	**Δ**	**CS**	**AR**	**CS**	**AR**	**CS**	**AR**	**CS**	**AR**
072	24	24	24	=	0.0	x	x	x	x	x	x	x	x
	18	24	30	+	0.2	x	x	x	x	x	x	x	x
	30	24	18	-	0.2	x	x	x	x	x	x	x	x
	12	24	36	+	0.4	x	x	x	x	x	x	x	x
	36	24	12	-	0.4	x	x	x	x	x	x	x	x
108	36	36	36	=	0.0	x	x	x	x	x	x	x	x
	27	36	45	+	0.2	x	x	x	x	x	x	x	x
	45	36	27	-	0.2	x	x	x	x	x	x	x	x
	18	36	54	+	0.4	x	x	x	x	x	x	x	x
	54	36	18	-	0.4	x	x	x	x	x	x	x	x
144	48	48	48	=	0.0	x	x	x	x	x	x	x	x
	36	48	60	+	0.2	x	x	x	x	x	x	x	x
	60	48	36	-	0.2	x	x	x	x	x	x	x	x
	24	48	72	+	0.4	x	x	x	x	x	x	x	x
	72	48	24	-	0.4	x	x	x	x	x	x	x	x

Note. CS = Compound symmetric; AR = First-order autoregressive; Δ = Coefficient of sample size variation; = stands for null pairing of unequal covariance matrices but equal group sizes; + stands for positive pairing of unequal covariance matrices and unequal group sizes; – stands for negative pairing of unequal covariance matrices and unequal group sizes.

Type I Error Rates for Tests of the Groups x Occasions Interaction Effect

Normally Distributed Data. Table 4 gives the empirical Type I error rates obtained in the simulation for the interaction effect when data were sampled from a multivariate normal distribution.

An inspection of the results in Table 4 indicates that, the *BF* statistic was able to control the Type I error rates across all of the investigated conditions, except for negative pairing condition, when N = 72, and Δ = .40. In this case, the same as it happened for the main effect tests and normally distributed data; the procedure had a tendency to have Type I error rates below the lower limit of Bradley's (1978) liberal criterion interval. However, the results in Table 4 show that the *WJ* procedure did not provide a robust test of the within-subjects interaction effect, given that exhibits poor control of the Type I error rates for many of the investigated conditions. A careful examination of the Table 4 reveals that, when there was an inverse relationship between sample sizes and dispersion matrices and Δ = .40, the *WJ* procedure was always liberal and error rates were, in some cases, may become severely inflated. In fact, in Table 4, it is readily seen that while *BF* statistic was able to control the Type I error rates in 110 of the 120 investigated conditions the *WJ* test had a liberal behavior in more than half of the examined conditions. Consistent with the findings of other researches, including Keselman and Lix (1997), the degree of liberalness of the *WJ* test decreasing as the sample sizes increases.

Nonnormally Distributed Data. Table 5 contains the empirical rates of Type I error for the interaction effect when data were sampled from a chi-squared distribution with three degrees of freedom.

As seen from the table 5, when the multivariate normality assumption was violated, the pattern of results associated with the *BF* statistic was very similar to the one observed when the normality assumption was satisfied. For this procedure, the impact of non-normality on Type I error rates is modest. With respect to the *WJ* procedure, error rates associated with the skewed distribution were almost always larger than those obtained for the normal distribution; in particular, for α = .01. For positive pairings, Type I error rates associated with the skewed distribution were not always

Table 2
Empirical Type I error rates for the within-subjects main effect and multivariate normal distribution

						Covariance Ratio							
						1/3: 1: 5/3				**1/5: 1: 9/5**			
						BF_M		WJ_M		BF_M		WJ_M	
N	**Struct.**	n_1	n_2	n_3	Δ	**α=.05**	**α=.01**	**α=.05**	**α=.01**	**α=.05**	**α=.01**	**α=.05**	**α=.01**
072	CS	24	24	24	0.0	0.0503	0.0094	0.0546	0.0103	0.0454	0.0090	0.0500	0.0100
072	CS	18	24	30	0.2	0.0464	0.0095	0.0490	0.0097	0.0468	0.0080	0.0486	0.0082
072	CS	30	24	18	0.2	0.0416	0.0060	0.0543	0.0088	0.0410	0.0074	0.0558	0.0107
072	CS	12	24	36	0.4	0.0515	0.0113	0.0541	0.0114	0.0515	0.0110	0.0534	0.0111
072	CS	36	24	12	0.4	**0.0151**	**0.0020**	0.0539	0.0071	**0.0223**	**0.0125**	0.0642	0.0078
072	AR	24	24	24	0.0	0.0481	0.0107	0.0530	0.0109	0.0479	0.0097	0.0546	0.0103
072	AR	18	24	30	0.2	0.0461	0.0087	0.0473	0.0089	0.0471	0.0108	0.0491	0.0108
072	AR	30	24	18	0.2	0.0418	0.0090	0.0514	0.0115	0.0411	0.0070	0.0543	0.0105
072	AR	12	24	36	0.4	0.0480	0.0084	0.0498	0.0085	0.0477	0.0090	0.0494	0.0092
072	AR	36	24	12	0.4	**0.0186**	**0.0011**	0.0588	0.0083	**0.0215**	**0.0200**	0.0610	0.0091
108	CS	36	36	36	0.0	0.0503	0.0090	0.0508	0.0086	0.0510	0.0094	0.0628	0.0099
108	CS	27	36	45	0.2	0.0510	0.0104	0.0509	0.0098	0.0505	0.0116	0.0597	0.0121
108	CS	45	36	27	0.2	0.0453	0.0074	0.0487	0.0076	0.0456	0.0081	0.0727	0.0117
108	CS	18	36	54	0.4	0.0464	0.0078	0.0465	0.0075	0.0488	0.0104	0.0555	0.0106
108	CS	54	36	18	0.4	0.0358	0.0058	0.0490	0.0069	0.0372	0.0054	0.0561	0.0105
108	AR	36	36	36	0.0	0.0464	0.0103	0.0471	0.0100	0.0450	0.0092	0.0553	0.0099
108	AR	27	36	45	0.2	0.0510	0.0107	0.0509	0.0104	0.0514	0.0109	0.0599	0.0105
108	AR	45	36	27	0.2	0.0458	0.0093	0.0494	0.0098	0.0454	0.0077	0.0711	0.0114
108	AR	18	36	54	0.4	0.0491	0.0113	0.0490	0.0108	0.0506	0.0115	0.0585	0.0119
108	AR	54	36	18	0.4	0.0339	0.0052	0.0498	0.0092	0.0394	0.0064	0.0564	0.0113
144	CS	48	48	48	0.0	0.0482	0.0107	0.0481	0.0103	0.0463	0.0089	0.0462	0.0086
144	CS	36	48	60	0.2	0.0513	0.0104	0.0504	0.0099	0.0481	0.0095	0.0476	0.0086
144	CS	60	48	36	0.2	0.0466	0.0089	0.0471	0.0087	0.0472	0.0091	0.0489	0.0089
144	CS	24	48	72	0.4	0.0539	0.0107	0.0532	0.0098	0.0485	0.0099	0.0477	0.0091
144	CS	72	48	24	0.4	0.0443	0.0068	0.0522	0.0087	0.0431	0.0059	0.0528	0.0082
144	AR	48	48	48	0.0	0.0461	0.0081	0.0460	0.0079	0.0470	0.0103	0.0475	0.0098
144	AR	36	48	60	0.2	0.0500	0.0093	0.0482	0.0085	0.0495	0.0097	0.0492	0.0092
144	AR	60	48	36	0.2	0.0459	0.0100	0.0464	0.0095	0.0488	0.0085	0.0505	0.0087
144	AR	24	48	72	0.4	0.0479	0.0103	0.0474	0.0101	0.0500	0.0087	0.0490	0.0081
144	AR	72	48	24	0.4	0.0406	0.0080	0.0491	0.0092	0.0413	0.0053	0.0494	0.0075

Note. BF_M = Brown-Forsythe main effect test; WJ_M = Welch-James main effect test; CS = Compound symmetric; AR = First-order autoregressive; Δ = Coefficient of sample size variation; Bold values are not contained in the interval $1/2\alpha \leq \hat{\alpha} \leq 3/2\alpha$.

larger than those obtained for the normal distribution. However, for balanced designs and negative pairings, error rates associated with the skewed distribution were always larger than those obtained for the normal distribution. In fact, for the *BF* test 8 conditions resulted in Type I error rates below the lower limit of Bradley's (1978) liberal criterion interval. Whereas, for the *WJ* test 85 conditions resulted in Type I error rates above upper limit of Bradley's liberal criterion.

As was true for the main effect and skewed data, covariance ratios and pattern of covariance matrices had a superior effect on the robustness of both procedures; especially, the degree of heterogeneity of the covariance matrices.

Finally, Table 6 gives a summary of the number of empirical Type I error rates above or below of interval $\alpha/2 \leq \hat{\alpha} \leq 3/2\alpha$. Each below and above cell corresponds to 10 conditions because the count is aggregated over the 2 alpha levels and 5 pairing conditions.

An inspection of the results in Table 6 indicates that, the *BF* statistic was able to control the Type I error rates in 432 of the 480 investigated conditions. In fact, for the *BF* test 28 conditions resulted in Type I error rates below the lower limit of Bradley's liberal criterion and 20 above upper limit of Bradley's liberal criterion. Whereas, for the *WJ* test 177 conditions resulted in Type I error rates above upper limit of Bradley's liberal criterion.

Discussion and conclusions

The purpose of this investigation was to compare the performance of the modified *BF* approach presented by Vallejo *et al.* (in press) with the performance of Johansen's (1980) solution, when testing within-subjects main and interaction effects in unbalanced multivariate split-plot designs. Specifically, we examined the robustness of these procedures when the homogeneity of the covariance matrices is not satisfied and data were obtained from the non-normal chi-squared distribution.

The results indicate that when covariance homogeneity assumption was violated, but the normality assumption is satisfied, both the *BF* and *WJ* test show a good control of Type I error rates across all of the investigated conditions for the within-subjects main effect. Although, for negative pairings and severe values of

Table 3
Empirical Type I error rates for the within-subjects main effect and multivariate nonnormal distribution

						Covariance Ratio							
						1/3: 1: 5/3				**1/5: 1: 9/5**			
						BF_M		WJ_M		BF_M		WJ_M	
N	**Struct.**	n_1	n_2	n_3	Δ	**α=.05**	**α=.01**	**α=.05**	**α=.01**	**α=.05**	**α=.01**	**α=.05**	**α=.01**
072	CS	24	24	24	0.0	0.0608	0.0122	0.0611	0.0126	0.0668	**0.0170**	0.0729	**0.0190**
072	CS	18	24	30	0.2	0.0590	0.0123	0.0612	0.0125	0.0629	0.0146	0.0654	**0.0151**
072	CS	30	24	18	0.2	0.0486	0.0101	0.0615	0.0133	0.0556	0.0113	**0.0757**	**0.0168**
072	CS	12	24	36	0.4	0.0610	0.0138	0.0628	0.0135	0.0662	**0.0159**	0.0682	**0.0157**
072	CS	36	24	12	0.4	0.0236	**0.0027**	**0.0756**	0.0137	0.0269	**0.0036**	**0.0827**	**0.0159**
072	AR	24	24	24	0.0	0.0548	0.0128	0.0604	0.0133	0.0682	**0.0195**	**0.0754**	**0.0175**
072	AR	18	24	30	0.2	0.0597	**0.0151**	0.0620	**0.0154**	0.0705	**0.0174**	0.0737	**0.0178**
072	AR	30	24	18	0.2	0.0520	0.0112	0.0670	0.0142	0.0605	**0.0159**	**0.0770**	**0.0201**
072	AR	12	24	36	0.4	0.0607	**0.0164**	0.0676	**0.0165**	0.0672	**0.0189**	0.0692	**0.0190**
072	AR	36	24	12	0.4	0.0248	**0.0032**	**0.0763**	0.0126	0.0271	**0.0440**	**0.0859**	**0.0176**
108	CS	36	36	36	0.0	0.0569	0.0125	0.0579	0.0126	0.0597	0.0132	0.0603	0.0128
108	CS	27	36	45	0.2	0.0580	0.0126	0.0579	0.0124	0.0604	0.0124	0.0605	0.0119
108	CS	45	36	27	0.2	0.0512	0.0112	0.0550	0.0117	0.0607	0.0145	0.0718	**0.0172**
108	CS	18	36	54	0.4	0.0545	0.0115	0.0544	0.0146	0.0621	0.0148	0.0622	0.0150
108	CS	54	36	18	0.4	0.0400	0.0077	0.0570	0.0130	0.0443	0.0084	0.0650	0.0145
108	AR	36	36	36	0.0	0.0547	0.0121	0.0553	0.0116	0.0636	**0.0170**	0.0654	**0.0166**
108	AR	27	36	45	0.2	0.0513	0.0117	0.0511	0.0108	0.0670	**0.0194**	0.0667	**0.0191**
108	AR	45	36	27	0.2	0.0565	0.0143	0.0604	**0.0151**	0.0691	**0.0176**	**0.0813**	**0.0205**
108	AR	18	36	54	0.4	0.0568	0.0141	0.0604	0.0135	0.0644	**0.0179**	0.0642	**0.0171**
108	AR	54	36	18	0.4	0.0465	0.0078	0.0663	0.0128	0.0486	0.0091	0.0717	**0.0167**
144	CS	48	48	48	0.0	0.0543	0.0132	0.0541	0.0127	0.0562	0.0126	0.0561	0.0119
144	CS	36	48	60	0.2	0.0624	0.0146	0.0604	0.0141	0.0619	0.0149	0.0614	0.0143
144	CS	60	48	36	0.2	0.0619	0.0142	0.0631	0.0143	0.0602	**0.0161**	0.0623	**0.0159**
144	CS	24	48	72	0.4	0.0626	**0.0152**	0.0618	0.0146	0.0624	0.0131	0.0613	0.0123
144	CS	72	48	24	0.4	0.0548	0.0115	0.0632	0.0135	0.0601	0.0128	0.0746	**0.0161**
144	AR	48	48	48	0.0	0.0555	0.0124	0.0553	0.0116	0.0620	0.0139	0.0619	0.0139
144	AR	36	48	60	0.2	0.0663	0.0148	0.0643	0.0143	0.0639	**0.0153**	0.0633	0.0148
144	AR	60	48	36	0.2	0.0601	0.0139	0.0611	0.0137	0.0615	**0.0154**	0.0627	**0.0153**
144	AR	24	48	72	0.4	0.0593	0.0128	0.0588	0.0125	0.0608	0.0147	0.0602	0.0142
144	AR	72	48	24	0.4	0.0581	0.0117	0.0670	0.0142	0.0611	0.0128	0.0749	**0.0161**

Note. BF_M = Brown-Forsythe main effect test; WJ_M = Welch-James main effect test; CS = Compound symmetric; AR = First-order autoregressive; Δ = Coefficient of sample size variation; Bold values are not contained in the interval $1/2\alpha \leq \hat{\alpha} \leq 3/2\alpha$.

coefficient of sample size variation, the *WJ* test seems preferable as a test of the within-subjects main effect, because it is never too conservative neither excessively liberal test.

When normality and homogeneity assumptions are jointly violated, the *WJ* test does not perform as well with those sample sizes that can be considered the norm, rather than the exception in the psychological and educational researches (see Kowalchuk *et al*, 1997). In this case, at least for the conditions included in our study, the *BF* approach is preferable a test of the within-subjects main effect. However, it is important to remember that, for the nonormal data, none of the procedures was able to control the rates of Type I error in all of the investigated conditions.

With regard to the test of the interaction effect, our results indicate that the *BF* procedure can effectively control the rate of Type I errors when group variance-covariance matrices are heterogeneous, even when the data were sampled from a chi-squared distribution with three degrees of freedom. This finding held even when the degree of heterogeneity of the covariance matrices was varied across the design. As with the main effect, the procedure tends to be conservative for negative pairings and severe values of coefficient of sample size variation. On the other hand, our results also indicate that when the interest lies in the interaction, the *WJ* test is not a adequate solution, since the sample sizes required to achieve robustness could be unreasonably large, particularly when the multivariate normality assumption is violated. For very large sample sizes the procedure appear to be robust. Nevertheless, sample sizes superiors to 200 subjects could be required. Unfortunately, according to a survey conducted by Kowalchuk *et a*l. (1996), these values are not frequent in the current educational and psychological investigation. This result is consistent with the findings of Algina and Keselman (1997) and Keselman and Lix (1997).

Consequently, because the *WJ* procedure require large sample sizes to obtain robust test of within-subjects effects in multivariate split-plot designs, in particular of the within-subjects interaction effects, when the multivariate normality and variance homogeneity assumptions are not satisfied, we recommended that researches use the *BF* procedure. In addition of the available results in Vallejo *et al*. (in press), this recommendation is based in that in a majority of the conditions used in the study the *BF* test was more

Table 4
Empirical Type I error rates for the interaction effect and multivariate normal distribution

						Covariance Ratio							
						1/3: 1: 5/3				**1/5: 1: 9/5**			
						BF_I		WJ_I		BF_I		WJ_I	
N	**Struct.**	$\mathbf{n_1}$	$\mathbf{n_2}$	$\mathbf{n_3}$	Δ	**α=.05**	**α=.01**	**α=.05**	**α=.01**	**α=.05**	**α=.01**	**α=.05**	**α=.01**
072	CS	24	24	24	0.0	0.0507	0.0116	**0.0811**	**0.0178**	0.0458	0.0084	**0.0803**	**0.0175**
072	CS	18	24	30	0.2	0.0514	0.0114	**0.0751**	**0.0167**	0.0518	0.0126	**0.0751**	**0.0185**
072	CS	30	24	18	0.2	0.0416	0.0071	**0.1038**	**0.0286**	0.0418	0.0080	**0.1200**	**0.0384**
072	CS	12	24	36	0.4	0.0544	0.0123	**0.0803**	**0.0186**	0.0539	0.0128	**0.0775**	**0.0195**
072	CS	36	24	12	0.4	**0.0191**	**0.0022**	**0.2500**	**0.1123**	**0.0181**	**0.0020**	**0.2961**	**0.1431**
072	AR	24	24	24	0.0	0.0494	0.0085	**0.0787**	**0.0188**	0.0460	0.0085	**0.0819**	**0.0183**
072	AR	18	24	30	0.2	0.0536	0.0107	**0.0755**	**0.0176**	0.0529	0.0120	**0.0794**	**0.0179**
072	AR	30	24	18	0.2	0.0389	0.0066	**0.1018**	**0.0285**	0.0399	0.0066	**0.1254**	**0.0376**
072	AR	12	24	36	0.4	0.0571	0.0119	**0.0857**	**0.0199**	0.0571	**0.0151**	**0.0768**	**0.0185**
072	AR	36	24	12	0.4	**0.0210**	**0.0034**	**0.2567**	**0.1132**	**0.0162**	**0.0018**	**0.2954**	**0.1510**
108	CS	36	36	36	0.0	0.0530	0.0125	0.0603	0.0111	0.0524	0.0128	0.0630	0.0133
108	CS	27	36	45	0.2	0.0548	0.0138	0.0599	0.0127	0.0584	0.0129	0.0607	0.0116
108	CS	45	36	27	0.2	0.0472	0.0094	0.0718	**0.0157**	0.0503	0.0097	0.0742	**0.0179**
108	CS	18	36	54	0.4	0.0594	0.0130	0.0634	0.0128	0.0578	0.0125	0.0583	0.0115
108	CS	54	36	18	0.4	0.0374	0.0063	**0.1136**	**0.0348**	0.0346	0.0056	**0.1397**	**0.0475**
108	AR	36	36	36	0.0	0.0528	0.0107	0.0615	0.0107	0.0549	0.0135	0.0623	0.0150
108	AR	27	36	45	0.2	0.0543	0.0140	0.0569	0.0138	0.0571	0.0142	0.0592	0.0139
108	AR	45	36	27	0.2	0.0469	0.0082	0.0666	**0.0153**	0.0483	0.0087	**0.0761**	**0.0174**
108	AR	18	36	54	0.4	0.0573	0.0133	0.0590	0.0129	0.0610	0.0150	0.0604	0.0115
108	AR	54	36	18	0.4	0.0358	0.0056	**0.1054**	**0.0325**	0.0343	0.0051	**0.1331**	**0.0484**
144	CS	48	48	48	0.0	0.0531	0.0109	0.0593	0.0131	0.0553	0.0144	0.0547	0.0107
144	CS	36	48	60	0.2	0.0538	0.0111	0.0541	0.0121	0.0560	0.0146	0.0531	0.0105
144	CS	60	48	36	0.2	0.0501	0.0101	0.0626	0.0139	0.0513	0.0131	0.0600	0.0126
144	CS	24	48	72	0.4	0.0597	0.0140	0.0564	0.0121	0.0640	**0.0169**	0.0573	0.0116
144	CS	72	48	24	0.4	0.0454	0.0092	**0.0850**	**0.0224**	0.0429	0.0083	**0.0896**	**0.0244**
144	AR	48	48	48	0.0	0.0576	0.0142	0.0590	0.0123	0.0571	0.0148	0.0541	0.0121
144	AR	36	48	60	0.2	0.0614	0.0138	0.0542	0.0103	0.0597	0.0147	0.0563	0.0109
144	AR	60	48	36	0.2	0.0502	0.0121	0.0619	0.0128	0.0527	0.0124	0.0583	0.0145
144	AR	24	48	72	0.4	0.0591	0.0134	0.0507	0.0114	0.0567	0.0150	0.0523	0.0123
144	AR	72	48	24	0.4	0.0447	0.0082	**0.0858**	**0.0235**	0.0418	0.0080	**0.0908**	**0.0252**

Note. BF_I = Brown-Forsythe interaction effect test; WJ_I = Welch-James interaction effect test; CS = Compound symmetric; AR = First-order autoregressive; Δ = Coefficient of sample size variation; Bold values are not contained in the interval $1/2\alpha \leq \hat{\alpha} \leq 3/2\alpha$.

robust than the *WJ* test. In short, the control of Type I error rates was achieved in 90 percent of the cases with *BF* test, and only in 63 percent of the cases with *WJ* test. Thus, in our opinion, applied researchers should be comfortable using the modified *BF* test to analyze multivariate repeated measures hypotheses when the assumptions of the general linear model are violated.

As final note, four lines of additional research can be of interest. First, it is very important to investigate whether the multivariate *BF* procedure offers robust tests when covariance matrices vary across groups but are not multiples of one another. Second, in the context of multivariate designs, it is not known whether the performance of the tests will change using trimmed means and Winsorized variances. However, the results obtained in the context univariate are encouraging (see, Wilcox, Keselman, Muska and Cribbie, 2000). Third, which of the robust procedures will be most sensitive for detecting treatment effects. Fourth, additional research manipulating other types of nonnormal distributions, both symmetric and asymmetric distributions with light tail and heavy tail, might also be investigated.

Acknowledgements

Esta investigación fue subvencionado por por una ayuda del Ministerio de Educación (DGICYT PS95-0228) y de la Universidad de Oviedo (IR99-0622).

Parte de los resultados del presente trabajo fueron presentados en el *Sixth European Congress of Psychology*, Julio de 1999, Roma.

Table 5
Empirical Type I error rates for the interaction effect and multivariate nonnormal distribution

						Covariance Ratio							
						1/3: 1: 5/3				**1/5: 1: 9/5**			
						BF_I		WJ_I		BF_I		WJ_I	
N	**Struct.**	n_1	n_2	n_3	Δ	**α=.05**	**α=.01**	**α=.05**	**α=.01**	**α=.05**	**α=.01**	**α=.05**	**α=.01**
072	CS	24	24	24	0.0	0.0478	0.0089	**0.0953**	**0.0217**	0.0484	0.0098	**0.1056**	**0.0308**
072	CS	18	24	30	0.2	0.0527	0.0093	**0.0808**	**0.0192**	0.0539	0.0114	**0.0913**	**0.0229**
072	CS	30	24	18	0.2	0.0400	0.0062	**0.1192**	**0.0358**	0.0451	0.0085	**0.1579**	**0.0546**
072	CS	12	24	36	0.4	0.0579	0.0146	**0.0871**	**0.0195**	0.0510	0.0120	**0.0821**	**0.0206**
072	CS	36	24	12	0.4	**0.0216**	**0.0023**	**0.2754**	**0.1265**	**0.0182**	**0.0017**	**0.3356**	**0.1677**
072	AR	24	24	24	0.0	0.0524	0.0090	**0.0982**	**0.0238**	0.0517	0.0105	**0.1189**	**0.0354**
072	AR	18	24	30	0.2	0.0504	0.0097	**0.0870**	**0.0195**	0.0581	0.0135	**0.0962**	**0.0253**
072	AR	30	24	18	0.2	0.0412	0.0074	**0.1369**	**0.0391**	0.0421	0.0077	**0.1646**	**0.0537**
072	AR	12	24	36	0.4	0.0510	0.0112	**0.0841**	**0.0172**	0.0524	0.0110	**0.0893**	**0.0216**
072	AR	36	24	12	0.4	**0.0224**	**0.0020**	**0.2769**	**0.1287**	**0.0189**	**0.0027**	**0.3444**	**0.1802**
108	CS	36	36	36	0.0	0.0542	0.0123	0.0724	**0.0179**	0.0596	0.0125	**0.0859**	**0.0219**
108	CS	27	36	45	0.2	0.0503	0.0100	0.0631	0.0124	0.0582	0.0128	0.0704	**0.0161**
108	CS	45	36	27	0.2	0.0516	0.0117	**0.0906**	**0.0239**	0.0502	0.0110	**0.1027**	**0.0289**
108	CS	18	36	54	0.4	0.0574	0.0123	0.0646	0.0131	0.0607	**0.0163**	0.0665	0.0143
108	CS	54	36	18	0.4	0.0352	0.0062	**0.1325**	**0.0456**	0.0357	0.0057	**0.1694**	**0.0636**
108	AR	36	36	36	0.0	0.0576	0.0129	**0.0773**	**0.0200**	0.0599	0.0135	**0.0871**	**0.0222**
108	AR	27	36	45	0.2	0.0547	0.0133	0.0698	0.0145	0.0577	0.0143	0.0718	**0.0162**
108	AR	45	36	27	0.2	0.0502	0.0102	**0.0915**	**0.0249**	0.0509	0.0100	**0.1018**	**0.0297**
108	AR	18	36	54	0.4	0.0569	0.0121	0.0638	0.0135	0.0537	0.0137	0.0668	**0.0153**
108	AR	54	36	18	0.4	0.0382	0.0078	**0.1437**	**0.0479**	0.0382	0.0060	**0.1849**	**0.0756**
144	CS	48	48	48	0.0	0.0554	0.0132	0.0656	0.0137	0.0582	0.0120	0.0699	**0.0162**
144	CS	36	48	60	0.2	0.0557	0.0126	0.0596	0.0125	0.0584	0.0120	0.0700	**0.0160**
144	CS	60	48	36	0.2	0.0514	0.0291	0.0732	**0.0180**	0.0589	0.0143	**0.0862**	**0.0228**
144	CS	24	48	72	0.4	0.0545	0.0117	0.0553	0.0096	0.0629	**0.0155**	0.0602	0.0113
144	CS	72	48	24	0.4	0.0474	0.0084	**0.1105**	**0.0338**	0.0479	0.0099	**0.1221**	**0.0415**
144	AR	48	48	48	0.0	0.0555	0.0127	0.0633	0.0146	0.0584	0.0141	0.0735	**0.0153**
144	AR	36	48	60	0.2	0.0662	0.0141	0.0669	0.0147	0.0589	0.0143	0.0619	**0.0151**
144	AR	60	48	36	0.2	0.0566	0.0126	**0.0826**	**0.0209**	0.0557	0.0132	**0.0850**	**0.0233**
144	AR	24	48	72	0.4	0.0571	0.0137	0.0592	0.0104	0.0648	0.0142	0.0618	0.0125
144	AR	72	48	24	0.4	0.0495	0.0091	**0.1189**	**0.0382**	0.0486	0.0101	**0.1312**	**0.0453**

Note. BF_I = Brown-Forsythe interaction effect test; WJ_I = Welch-James interaction effect test; CS = Compound symmetric; AR = First-order autoregressive; Δ = Coefficient of sample size variation; Bold values are not contained in the interval $1/2\alpha \leq \hat{\alpha} \leq 3/2\alpha$.

Table 6
Number of empirical alpha levels above and below the nominal alpha level by the interval $.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$

			Normal				Nonnormal			
Sample	**Covariance**	**Covariance**	**BF**		**WJ**		**BF**		**WJ**	
Sizes	**Structures**	**Ratios**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**
Main effect										
072	CS	1/3:1: 5/3	2	0	0	0	1	0	0	1
072	CS	1/5:1: 9/5	2	0	0	0	1	2	0	7
072	AR	1/3:1: 5/3	2	0	0	0	1	2	0	3
072	AR	1/5:1: 9/5	2	0	0	0	1	4	0	8
108	CS	1/3:1: 5/3	0	0	0	0	0	0	0	0
108	CS	1/5:1: 9/5	0	0	0	0	0	0	0	1
108	AR	1/3:1: 5/3	0	0	0	0	0	0	0	1
108	AR	1/5:1: 9/5	0	0	0	0	0	4	0	6
144	CS	1/3:1: 5/3	0	0	0	0	0	1	0	0
144	CS	1/5:1: 9/5	0	0	0	0	0	1	0	2
144	AR	1/3:1: 5/3	0	0	0	0	0	0	0	0
144	AR	1/5:1: 9/5	0	0	0	0	0	2	0	2
Subtotal			*8*	*0*	*0*	*0*	*4*	*16*	*0*	*31*
Interaction effect										
072	CS	1/3:1: 5/3	2	0	0	10	2	0	0	10
072	CS	1/5:1: 9/5	2	0	0	10	2	0	0	10
072	AR	1/3:1: 5/3	2	0	0	10	2	0	0	10
072	AR	1/5:1: 9/5	2	1	0	10	2	0	0	10
108	CS	1/3:1: 5/3	0	0	0	3	0	0	0	5
108	CS	1/5:1: 9/5	0	0	0	3	0	1	0	7
108	AR	1/3:1: 5/3	0	0	0	3	0	0	0	6
108	AR	1/5:1: 9/5	0	0	0	4	0	0	0	8
144	CS	1/3:1: 5/3	0	0	0	2	0	0	0	3
144	CS	1/5:1: 9/5	0	1	0	2	0	1	0	6
144	AR	1/3:1: 5/3	0	0	0	2	0	0	0	4
144	AR	1/5:1: 9/5	0	0	0	2	0	0	0	6
Subtotal			*8*	*2*	*0*	*61*	*8*	*2*	*0*	*85*
TOTAL			**16**	**2**	**0**	**61**	**12**	**18**	**0**	**116**

Note. BF_I = Brown-Forsy test; WJ_I = Welch-James test; A= level above the nominal alpha; B = level below the nominal alpha; CS = Compound symmetric; AR = First-order autoregressive; Bold values are not contained in the interval $1/2\alpha \leq \hat{\alpha} \leq 3/2\alpha$.

References

Algina, J., & Keselman, H. J. (1997). Testing repeated measures hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the Welch-James test. *Multivariate Behavioral Research, 32,* 255-274.

Algina, J., & Oshima, T. C. (1995). An improved general approximation test for the main effect in a split-plot design. *British Journal of Mathematical and Statistical Psychology, 48,* 149-160.

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

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

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.

Davidson, M. L. (1972). Univariate versus multivariate test in repeated measures experiments. *Psychological Bulletin, 77*, 446-452.

GAUSS (1997). *The Gauss System* (Vers. 3.2.32). Washington: Aptech Systems, Inc.

Hastings, N. A. J., & Peacock, J. B. (1975). *Statistical Distributions: A Handbook for Students and Practitioners.* New York: John Wiley.

Huynh, H. (1978). Some approximate tests for repeated measurement designs. *Psychometrika, 43*, 161-165.

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

Jones, R. H. (1993). *Longitudinal Data with Serial Correlation: A State-Space Approach.* London: Chapman and Hall.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1999). A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology, 52,* 63-78.

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

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

Keselman, H. J., & Lix, L. M. (1997). Analysing multivariate repeated measures designs when covariance matrices are heterogeneous. *British Journal of Mathematical and Statistical Psychology, 50*, 319-338.
Khatri, C. G. (1980). Quadratic forms in normal variables. In P. R. Krisnaiah (Ed.). *Handbook of Statistics 1: Analysis of Variance*. New York: North Holland Publishing Company.
Kinderman, A. J., & Ramage, J. G. (1976). Computer generation of normal random numbers. *Journal of the American Statistical Association, 77*, 893-896.
Kowalchuk, J. C., Lix, L. M., & Keselman, H. J. (1996). The analysis of repeated measures designs. Paper presented at the Annual Meeting of the Psychometric Society, Banff, Alberta.
Mehrotra, D. V. (1997). Improving the Brown-Forsythe solution to the generalized Behrens-Fisher problem. *Communication in Statistics-Simulation and Computation, 26*, 1139-1145.
Mendoza, J. H., Toothaker, L. E., & Nicewander, W. J. (1974). A Monte Carlo comparison of the univariate and multivariate methods for the groups by trials repeated measures design. *Multivariate Behavioral Research, 9*, 165-178.
Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105, 156-166.*
Nel, D. G. & van der Merwe, C. A. (1986). A solution to the multivariate Behrens-Fisher problem. *Communications in Statistics-Theory and Methods, 15*, 3719-3735.
Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association, 69*, 894-908.
Rao, C. R. (1951). An asymptotic expansion of the distribution of Wilks's criterion. *Bulletin of the International Statistical Institute, 33,* Part 2, 177-180.
Rogan, C. J., Keselman, H. J., & Mendoza, J. L. (1979). Analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology, 32*, 269-286.
Rubin, A. S. (1983). The use of weighted contrast in analysis of models with heterogeneity of variance. *Proceedings of the Business and Economic Statistics Section, American Statistical Association,* 347-352.
SAS Institute (1996). *SAS/STAT Software: Changes and Enhancements through Release 6.11*. Cary, NC: SAS Institute Inc., charter, 18.
Scheffé, H. (1956). A mixed model for the analysis of variance. *Annals of Mathematical Statistics, 27*, 23-36.
Shoukri, M. M., & Pause, C. A. (1999). *Statistical Methods for Health Sciences (2nd ed.)*. New York: CRC Press.
Schuer, E. M., & Stoller, D. S. (1966). On the generation of normal random vectors. *Technometrics, 4*, 278-280.
Vallejo, G., & Menéndez, I. A. (1997). Una comparación de enfoques alternativos para el análisis de diseños multivariados de medidas repetidas. *Psicothema, 9*, 647-656.
Vallejo, G., & Escudero, J. R. (1998). Algunas soluciones aproximadas para diseños split-plot con matrices de covarianza arbitrarias. *Qüestiió: Quaderns d'Estadística i Investigació Operativa,22*, 463-468.
Vallejo, G., Fernández, P., Fidalgo, A. M., & Escudero, J. R. (1999). Comparación de la robustez de cuatro pruebas en un diseño multivariado split-plot. *Metodología de Investigación de las Ciencias del Comportamiento, 1*, 1-23.
Vallejo, G., Fidalgo, A. M., & Fernández, P. (1998). Efectos de la no esfericidad en el análisis de diseños multivariados de medidas repetidas. *Anales de Psicología, 14*, 249-268.
Vallejo, G., Fidalgo, A. M., & Fernández, P. (in press). Effects of covariance heterogeneity on three procedures for analysing multivariate repeated measures designs. *Multivariate Behavioral Research*.
Wilcox, R. R. (1989). Adjusting for unequal variances when comparing means in one-way and two-way fixed effects ANOVA models. *Journal of Educational Statistics, 14*, 269-278.
Wilcox, R. R., Keselman, H. J., Muska, J., & Cribbie, R. (2000). Repeated measures ANOVA: Some new results on comparing trimmed means and means. *British Journal of Mathematical and Statistical Psychology, 53*, 69-82.
Wilks, S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, *24*, 471-494.

Aceptado el 20 de junio de 2000