

Robust tests for multivariate factorial designs under heteroscedasticity

Guillermo Vallejo · Manuel Ato

Published online: 13 October 2011
© Psychonomic Society, Inc. 2011

Abstract The question of how to analyze several multivariate normal mean vectors when normality and covariance homogeneity assumptions are violated is considered in this article. For the two-way MANOVA layout, we address this problem adapting results presented by Brunner, Dette, and Munk (BDM; 1997) and Vallejo and Ato (modified Brown–Forsythe [MBF]; 2006) in the context of univariate factorial and split-plot designs and a multivariate version of the linear model (MLM) to accommodate heterogeneous data. Furthermore, we compare these procedures with the Welch–James (WJ) approximate degrees of freedom multivariate statistics based on ordinary least squares via Monte Carlo simulation. Our numerical studies show that of the methods evaluated, only the modified versions of the BDM and MBF procedures were robust to violations of underlying assumptions. The MLM approach was only occasionally liberal, and then by only a small amount, whereas the WJ procedure was often liberal if the interactive effects were involved in the design, particularly when the number of dependent variables increased and total sample size was small. On the other hand, it was also found that the MLM procedure was uniformly more powerful than its most direct competitors. The overall success rate was 22.4% for the BDM, 36.3% for the MBF, and 45.0% for the MLM.

Keywords MANOVA · Assumption violations · Empirical generalized least squares · Moments estimators · Ordinary least squares

The multivariate analysis of variance (MANOVA) is a natural generalization of the univariate analysis of variance (ANOVA) to multidimensional observations. That is, two or more possibly correlated dependent variables are simultaneously analyzed using the known values of one, two, or more factors. There are several reasons why a researcher should consider employing a MANOVA instead of separate ANOVAs for each dependent variable. First, by using a MANOVA, researchers can obtain a more detailed description of the phenomenon under investigation, and we also have a better chance of discovering the overall impact of the treatment effect. Second, by using a MANOVA, the researchers control the overall alpha level at the desired level. Third, by using a MANOVA, the researchers can increase the statistical power. Although the MANOVA offers some useful advantages, there are some disadvantages as well. One is that a MANOVA is harder to interpret than an ANOVA. Another is the degree of freedom (*df*) lost. For this reason, if the hypotheses of interest are univariate in nature, it may be desirable to conduct separate ANOVAs. However, in practice, researchers often find that the response variables are correlated (Farivar, Cunningham, & Hays, 2007).

Many psychological processes are too complex to be measured by a single response variable. For this reason, the MANOVA has been one of the most utilized multivariate methods in published behavioral research. In fact, the technique has been applied successfully in such diverse psychological areas as basic psychological processes (Lindemann & Bekkering, 2009), clinical psychology

G. Vallejo (✉)
Department of Psychology, University of Oviedo,
Plaza de Benito Feijóo, s/n,
33003 Oviedo, Spain
e-mail: gvallejo@uniovi.es

M. Ato
Department of Behavioral Science Methods,
University of Murcia,
Edificio Luis Vives,
30080 Espinardo Murcia, Spain
e-mail: matogar@um.es

(Crowell et al., 2008), developmental psychology (Gathercole, Pickering, Ambridge, & Wearing, 2004), educational psychology (Blöte, Van der Burg, & Klein, 2001), neuropsychology (Holtzer et al., 2007), and personality and social psychology (Jonas & Sassenberg, 2006). However, as has been pointed out in numerous multivariate texts, including Anderson (2003), Johnson and Wichern (2008), and Hair, Black, Babin, and Anderson (2010), in order for the MANOVA to work appropriately, the statistical hypotheses must be tested on the assumption that the observations are independently and normally distributed with a common covariance matrix.

For testing linear hypotheses in a multivariate design setup, the most often used test statistics are Lawley–Hotelling's trace, Pillai–Bartlett's trace, and Wilks's lambda criterion. Several studies have been conducted to investigate the extent to which MANOVA tests are robust to violations of underlying assumptions (Finch & Davenport, 2009; Hakstian, Roed, & Lind, 1979; Ito, 1980; Mardia, 1971; Olson, 1974; Sheehan-Holt, 1998). The general conclusion to be drawn from these studies is that for equal group sizes, violations of normality and equality of covariance matrices produce slightly distorted error rates. However, empirical findings also indicate that for unequal group sizes, MANOVA criteria result in biased (i.e., conservative or liberal) tests when assumptions are not met, particularly when homoscedasticity is violated. Therefore, before employing these test criteria, it is important to check the tenability of the MANOVA assumptions. Naik (2003) and Jamshidian and Schott (2007) provide a review of some of the common procedures adopted in practice. In spite of the fact that studies of robustness properties of test criteria have been confined to the one-way MANOVA, it seems reasonable to extend these results to the more complex multifactor MANOVA, taking into account what has been found in univariate settings (e.g., Vallejo, Fernández, & Livacic-Rojas, 2010a).

In order to avoid the negative impact that the heterogeneity of covariance matrices has on multivariate test criteria, several researchers have developed generalizations of the univariate solutions to the Behrens–Fisher problem for testing differences in mean vectors. Among them, Johansen (1980) generalized Welch's univariate approximate df solution to analyze multivariate data collected in one-way and multiway layouts. Coombs and Algina (1996) extended the univariate Brown–Forsythe (BF) approach to analyze multivariate data collected in a one-way layout, and Vallejo and Ato (2006) and Vallejo, Arnau, and Ato (2007) extended the approximate df BF procedure to the univariate and multivariate repeated measures context. Algina (1994) had previously presented its applications in the context of a split-plot design. Brunner, Dette, and Munk (1997) developed a robust solution for the analysis of univariate

heteroscedastic factorial designs based on Box's (1954) method of matching moments. A rank version of this approximation was used for Brunner, Munzel, and Puri (2002) on the nonparametric Behrens–Fisher problem in the multivariate context. A completely different approach for dealing with unequal covariance matrices is based on the concept of generalized p values proposed by Gamage, Mathew, and Weerahandi (2004). Another option for handling heteroscedastic data directly is to use bootstrapping and permutation resampling methods.

Finch and Davenport (2009) recently examined the performance of the approximate F MANOVA test criteria listed above with their Monte Carlo permutation test analogs, as implemented through the SAS Institute's (2008) PROC GLM program, when the underlying assumptions were not met. The results indicated that both the approximate F and Monte Carlo tests were liberal when normality and homogeneity covariance assumptions were simultaneously violated, and, in some cases, they were substantially liberal. More recently, Krishnamoorthy and Lu (2010) compared, under normality, Johansen's test, the generalized inference procedure proposed by Gamage et al. (2004), and the parametric bootstrap test. They concluded that the latter approach is the only test that appears to be satisfactory for all of the conditions considered. The Johansen procedure performs satisfactorily when the sample sizes are moderate, while the generalized inference procedure appears to be liberal even when the sample sizes are moderately large. Therefore, the parametric bootstrap test could be applied in situations where the Johansen test cannot. However, the computational effort is considerable, and under some conditions of departures from normality, the parametric bootstrap test is not always robust.

Vallejo, Fernández, and Livacic-Rojas (2010a) investigated the joint impact of violating variance homogeneity and normality on the performance of the Brunner et al. (1997) and Johansen (1980) procedures for analyzing univariate factorial designs. The results indicated that both approaches consistently controlled the rates of error when the shape of the distribution was symmetric. However, these methods can be affected by asymmetric distributions. Later, Vallejo, Ato, and Fernández, (2010b) demonstrated that it is possible to achieve robustness, even when data are extremely heterogeneous and severely skewed, applying the original Brunner et al. (1997) approach in conjunction with Hall's transformation. Previously, Vallejo and Ato (2006) had corrected the multivariate BF approach in a manner similar to that employed by Mehrotra (1997) to correct the univariate BF approach. In this work, Vallejo and Ato found that the modified BF approach (referred to hereafter as MBF) was typically robust for the conditions when the BF approach was not robust. The results in Brunner et al. (1997) and Vallejo and Ato, presented in the context of

univariate factorial designs or in the context of multivariate analysis of data collected in split-plot designs, can also be used for the problem of comparing several multivariate normal mean vectors when the covariance matrices are arbitrary.

Accordingly, the purpose of the present study was twofold: (1) to show how the Bruner et al. (1997) and Vallejo and Ato (2006) methods can be adapted to analyze multivariate data collected in two-way layouts and (2) to examine the operating characteristics (i.e., Type I rates and Type II error rates) of the new approximate df solutions for testing omnibus main and interaction effects when normality and covariance homogeneity assumptions are violated. In addition to examining the behavior of these procedures, for comparative purposes, Johansen's (1980) test and extensions of the linear model to accommodate multivariate heterogeneous data as implemented through the SAS PROC MIXED module were also studied. The former may serve as a benchmark, since it has been found to be generally robust under conditions similar to those investigated in our article. An advantage of the PROC MIXED approach, however, is that it can easily accommodate incomplete data and will tend to produce correct analyses provided the data are missing at random and the distributional assumptions are met. Also, the usual MANOVA test using Wilks's lambda criterion available in SAS PROC GLM is included, in spite of its well-known conservatism or liberalism, because the users of factorial designs almost exclusively demand this approach. In all cases, nonorthogonal two-way MANOVA designs, based on the unweighted multivariate linear model, were used.

Definition of the statistical tests

Consider a multivariate two-way layout where factor A has $j = 1, \dots, a$ levels and factor B has $k = 1, \dots, b$ levels with $i = 1, \dots, n_{jk}$ subjects per cell (j, k). In terms of the full rank (FR) or cell means model, the response vector $\mathbf{y}_{ijk} = (Y_{ijk1}, \dots, Y_{ijkp})'$ for the i th subject within the jk th cell of the factorial design is modeled by $\mathbf{y}_{ijk} = \boldsymbol{\mu}_{jk} + \mathbf{e}_{ijk}$, where $\boldsymbol{\mu}_{jk}$ represents the population means of the jk th cell and \mathbf{e}_{ijk} are independent random vectors of errors, each having a p -variate normal distribution with $E(\mathbf{e}_{ijk}) = \mathbf{0}$ and $V(\mathbf{e}_{ijk}) = \boldsymbol{\Sigma}_{jk}$. In some cases, the response vector \mathbf{y}_{ijk} is modeled by $\mathbf{y}_{ijk} = \boldsymbol{\mu} + \boldsymbol{\alpha}_j + \boldsymbol{\beta}_k + \boldsymbol{\gamma}_{jk} + \mathbf{e}_{ijk}$, where the parameter vector $\boldsymbol{\mu}$ is the overall mean, $\boldsymbol{\alpha}_j$ ($\boldsymbol{\beta}_k$) is the effect of the j th (k th) level of A (B), and $\boldsymbol{\gamma}_{jk}$ is the joint effect of treatment levels j and k . The link between the FR model and the less full rank (LFR) model is $\boldsymbol{\mu}_{jk} = \boldsymbol{\mu} + \boldsymbol{\alpha}_j + \boldsymbol{\beta}_k + \boldsymbol{\gamma}_{jk}$, with the constraints $\sum_{j=1}^a \boldsymbol{\alpha}_j = \sum_{k=1}^b \boldsymbol{\beta}_k = \sum_{j=1}^a \boldsymbol{\gamma}_{jk} = \sum_{k=1}^b \boldsymbol{\gamma}_{jk} = \mathbf{0}$, since the LFR model is overparameterized (i.e., the number of cell means is

less than the number of model parameters to be estimated).

By stacking the subvectors $\mathbf{y}_{111}, \dots, \mathbf{y}_{nab}$, both models can be considered in a unified framework under a multivariate linear model

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \quad (1)$$

where \mathbf{Y} is an $n \times p$ matrix of observed data, \mathbf{X} is an $n \times q$ fixed design matrix with rank $R = R(\mathbf{X}) = ab < q$ ($1 + a + b + ab$) for the LFR model and $R(\mathbf{X}) = ab = q$ for the FR model, \mathbf{B} is a $q \times p$ matrix that contains the unknown fixed effects common to all participants, and \mathbf{E} is an $n \times p$ matrix of random errors. We assume that the rows of \mathbf{Y} are normally and independently distributed within each of the ab cells or treatment combinations, with mean vector $\boldsymbol{\mu}_{jk}$ and covariance matrix $\boldsymbol{\Sigma}_{jk}$. The unbiased estimators of $\boldsymbol{\Sigma}_{jk}$ are $\widehat{\boldsymbol{\Sigma}}_{jk} = 1/(n_{jk} - 1) \mathbf{E}_{jk}$, where $\mathbf{E}_{jk} = \mathbf{Y}'_{jk} \mathbf{Y}_{jk} - \widehat{\boldsymbol{\beta}}'_{jk} \mathbf{X}'_{jk} \mathbf{Y}_{jk}$ are distributed independently as Wishart $W_p(n_{jk} - 1, \boldsymbol{\Sigma}_{jk})$ and $\widehat{\boldsymbol{\beta}}_{jk}$ is the ordinary least squares estimator of vector $\boldsymbol{\beta}_{jk}$ (Nel, 1997). We also assume that $n_{jk} - 1 \geq p$, so that $\widehat{\boldsymbol{\Sigma}}_{jk}^{-1}$ exists with a probability of one.

From an FR model perspective, the general linear hypothesis can be written using matrix notation as follows:

$$H_0(t): \mathbf{R}_t \boldsymbol{\mu} = \mathbf{0}, \quad (t = A, B, \text{ or } AB), \quad (2)$$

where \mathbf{R}_t is the contrast matrix associated with the specific hypothesis, $\boldsymbol{\mu} = (\boldsymbol{\mu}_{11}, \dots, \boldsymbol{\mu}_{1b}, \boldsymbol{\mu}_{21}, \dots, \boldsymbol{\mu}_{ab})'$ is a vector column of population cell means, and $\mathbf{0}$ is the null vector. Throughout the article, the matrices \mathbf{R}_t to test the multivariate null hypotheses $H_0(A)$, $H_0(B)$ and $H_0(AB)$ are defined as $\mathbf{R}_A = \mathbf{C}_A \otimes \mathbf{I}_p$, $\mathbf{R}_B = \mathbf{C}_B \otimes \mathbf{I}_p$, and $\mathbf{R}_{AB} = \mathbf{C}_{AB} \otimes \mathbf{I}_p$, where $\mathbf{C}_A = [((\mathbf{1}_{a-1} | - \mathbf{I}_{a-1}) \otimes (1/b) \mathbf{1}'_b)]$, $\mathbf{C}_B = [(1/a) \mathbf{1}'_a \otimes (\mathbf{1}_{b-1} | - \mathbf{I}_{b-1})]$, and $\mathbf{C}_{AB} = [(\mathbf{1}_{a-1} | - \mathbf{I}_{a-1}) \otimes (\mathbf{1}_{b-1} | - \mathbf{I}_{b-1})]$, are matrices of between-subjects contrasts with full row rank, $\mathbf{1}_{\alpha(b)}$ is an a (b)-dimensional vector of one, $\mathbf{I}_{\alpha(b)}$ is an a (b)-dimensional identity matrix, the symbol $|$ represents the augmented matrix obtained by appending the columns of two given matrices ($\cdot | \cdot$), and the circle times operator (\otimes) denotes the direct product of the matrices. Accordingly, \mathbf{R}_t forms a set of linearly independent contrasts among the levels of factors for each dependent variable. Similarly, the contrast vectors for conducting multivariate multiple comparisons involving both marginal means and interaction or tetrad contrasts are defined as $\mathbf{r}_A = (\mathbf{p}_{jj'} \otimes \mathbf{1}'_k) \otimes \mathbf{I}_p$, $\mathbf{r}_B = (\mathbf{1}'_j \otimes \mathbf{p}_{kk'}) \otimes \mathbf{I}_p$, and $\mathbf{r}_{AB} = (\mathbf{p}_{jj'} \otimes \mathbf{p}_{kk'}) \otimes \mathbf{I}_p$, where $\mathbf{P}_{jj'}$ contains the coefficients that contrast the j th and j' th row means and $\mathbf{p}_{kk'}$ contains the coefficients that contrast the k th and k' th column means.

Multivariate version of the Welch–James type statistic (WJ)

It is well known in the literature (see, e.g., Maxwell & Delaney, 2004) that the usual ANOVA can become very sensitive to departures from the homogeneity assumption, particularly in the unbalanced case. To overcome this problem, Welch (1951) and James (1951) suggested weighting the terms appearing in the F ratio to account for the different population variances and estimating the denominator df from the data. Johansen (1980) extended the Welch–James null distribution results to the regression setting. According to Lix, Algina, and Keselman (2003), the multivariate WJ-type statistic given by Johansen is

$$T_{WJ_t} = (\mathbf{R}_t \hat{\boldsymbol{\mu}})' (\mathbf{R}_t \hat{\boldsymbol{\Omega}} \mathbf{C} \mathbf{R}_t')^{-1} (\mathbf{R}_t \hat{\boldsymbol{\mu}}) / c, \tag{3}$$

where $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_{11} \dots \hat{\boldsymbol{\mu}}_{1b}, \hat{\boldsymbol{\mu}}_{21} \dots \hat{\boldsymbol{\mu}}_{ab})'$, $\hat{\boldsymbol{\Omega}} = \text{diag} (\hat{\boldsymbol{\Sigma}}_{11} / n_{11} \dots \hat{\boldsymbol{\Sigma}}_{ab} / n_{ab})$, $c = R(\mathbf{R}_t) + 2A - (6A) / [R(\mathbf{R}_t) + 2]$, and $A = \frac{1}{2} \sum_{jk} \left[\text{tr} \left(\hat{\boldsymbol{\Omega}} \mathbf{R}_t' (\mathbf{R}_t \hat{\boldsymbol{\Omega}} \mathbf{R}_t')^{-1} \mathbf{R}_t \mathbf{Q}_{jk} \right)^2 + \text{tr} \left(\hat{\boldsymbol{\Omega}} \mathbf{R}_t' (\mathbf{R}_t \hat{\boldsymbol{\Omega}} \mathbf{R}_t')^{-1} \mathbf{R}_t \mathbf{Q}_{jk} \right)^2 \right] / (n_{jk} - 1)$. The matrix \mathbf{Q}_{jk} is a block diagonal matrix of dimension abp associated with \mathbf{X}_{jk} , such that the (q, r) th block of \mathbf{Q}_{jk} is $\mathbf{I}_p \times p$ if $q = r = jk$ and is zero otherwise and $\text{tr}(\cdot)$ denotes the trace of a square matrix.

Johansen (1980) showed that, under $H_0(t): \mathbf{R}_t \boldsymbol{\mu} = \mathbf{0}$, T_{WJ_t} is distributed approximately as an F random variable with numerator $df \hat{f}_1 = R(\mathbf{R}_t)$ and denominator $df \hat{f}_2 = \hat{f}_1 (\hat{f}_1 + 2) / 3A$. Thus, $H_0(t)$ s are rejected if $T_{WJ_t} \geq F_{1-\alpha}(\hat{f}_1, \hat{f}_2)$.

Multivariate version of the ANOVA-type statistic (BDM)

For the analysis of univariate heteroscedastic factorial designs, Bruner et al. (1997) proposed a modification of the quadratic forms used in the usual ANOVA with estimated df based on a classical moment-matching approach by Box (1954). In this section, we extend the ANOVA-type statistic to the multivariate heteroscedastic factorial designs case. The corresponding multivariate version of the ANOVA-type statistic is given by

$$F_{N_t} = \frac{N \hat{\boldsymbol{\mu}}' \mathbf{P}_t \hat{\boldsymbol{\mu}}}{\text{tr}(\mathbf{P}_t^* \hat{\mathbf{V}})} = \frac{\hat{\mathbf{Q}}_1}{\hat{\mathbf{Q}}_2}, \tag{4}$$

where $N = \sum_{j=1}^a \sum_{k=1}^b n_{jk}$ is the total sample size, $\mathbf{P}_t = \mathbf{R}_t' (\mathbf{R}_t \mathbf{R}_t')^{-1} \mathbf{R}_t$ denotes the projection matrix on the image \mathbf{R}_t , \mathbf{P}_t^* denotes the diagonal matrix of the elements of \mathbf{P}_t , and $\hat{\mathbf{V}} = N \hat{\boldsymbol{\Omega}}$.

The distribution of the ANOVA-type statistic F_N can be approximated by a χ^2 distribution or by an F distribution with estimated df . For small samples sizes, Brunner et al. (1997) suggested approximating the distribution of F_N using the central F distribution with estimated numerator $df f_1$ and denominator $df f_2$. This is equivalent to using a ratio of two independent chi-square variables, each divided by its df —that is, $F_N \approx (\chi_{f_1}^2 / f_1) / (\chi_{f_2}^2 / f_2) = F(f_1, f_2)$.

Using results from Bruner et al. (1997) and using well-known theorems on the distribution of quadratic forms (see, e.g., Box, 1954; Mathai & Provost, 1992), it can be demonstrated that under $\mathbf{R}_t \boldsymbol{\mu} = \mathbf{0}$, the distribution of quadratic form $\mathbf{Q}_1 = N \boldsymbol{\mu} \mathbf{P}_t \boldsymbol{\mu}$ in Eq. 4 is asymptotically equivalent to quantity $Z = \sum_{j=1}^a \sum_{k=1}^b \sum_{l=1}^p \lambda_{jk}^{(p)} Z_{jk}^{(p)}$, where the Z s are mutually independently random variables, each distributed as chi-square with one df , and the λ s are the real nonzero latent roots of the matrix $\mathbf{P}_t \mathbf{V}$. The above sum is approximately distributed as $c_1 \chi_{f_1}^2$, where c_1 is a constant and $\chi_{f_1}^2$ is a central chi-square with f_1 df , and c_1 and f_1 determined in such a way that the first two moments (i.e., expectation and dispersion) of Z are equal to those of $c_1 \chi_{f_1}^2$, respectively. In particular, the numerator df is obtained by solving simultaneously the following equations:

$$\begin{aligned} E \left(\sum_{j=1}^a \sum_{k=1}^b \sum_{l=1}^p \lambda_{jk}^{(p)} Z_{jk}^{(p)} \right) &= \sum_{j=1}^a \sum_{k=1}^b \sum_{l=1}^p \lambda_{jk}^{(p)} \\ &= E(c_1 \chi_{f_1}^2) = c_1 f_1 \\ V \left(\sum_{j=1}^a \sum_{k=1}^b \sum_{l=1}^p \lambda_{jk}^{(p)} Z_{jk}^{(p)} \right) &= 2 \sum_{j=1}^a \sum_{k=1}^b \sum_{l=1}^p \lambda_{jk}^{(p)2} \\ &= V(c_1 \chi_{f_1}^2) = 2c_1^2 f \end{aligned} \tag{5}$$

—that is,

$$f_1 = \frac{\left(\sum_{j=1}^a \sum_{k=1}^b \sum_{l=1}^p \lambda_{jk}^{(p)2} \right)^2}{\left(\sum_{j=1}^a \sum_{k=1}^b \sum_{l=1}^p \lambda_{jk}^{(p)} \right)^2}. \tag{6}$$

Thus, it follows that $\mathbf{Q}_1 / c_1 f_1 \approx c_1 \chi_{f_1}^2 / c_1 f_1 \approx \chi_{f_1}^2 / f_1$. If \mathbf{V} was known, f_1 could easily be computed using $\sum_{j=1}^a \sum_{k=1}^b \sum_{l=1}^p \lambda_{jk}^{(p)} = \text{tr}(\mathbf{P}_t \hat{\mathbf{V}})$ and $\sum_{j=1}^a \sum_{k=1}^b \sum_{l=1}^p \lambda_{jk}^{(p)2} = \text{tr}(\mathbf{P}_t \hat{\mathbf{V}} \mathbf{P}_t \hat{\mathbf{V}})$.

Again applying the moment-matching principle, it is also possible to determine the estimated denominator df . Now let $\mathbf{Q}_2 = \text{tr}(\mathbf{P}_t^* \mathbf{V})$, where

$$\text{tr}(\mathbf{P}_t^* \mathbf{V}) = N \sum_{j=1}^a \sum_{k=1}^b \frac{1}{n_{jk}(n_{jk} - 1)} \sum_{l=1}^{n_{jk}} (\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{jk})' \mathbf{P}_{ll}^* (\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{jk}),$$

and \mathbf{P}_{ll}^* is the jk th p -dimensional submatrix of $\mathbf{P}_t^* = \{\mathbf{P}_{ll}^*\}$. The distribution of $\mathbf{Q}_2 = \text{tr}(\mathbf{P}_t^* \mathbf{V})$ is approximated by the distribution of $c_2 \chi_{f_2}^2 / f_2$ such that the first two moments coincide; that is,

$$E(\text{tr}(\mathbf{Q}_2)) = N \sum_{j=1}^a \sum_{k=1}^b \text{tr}(\mathbf{P}_{ll}^* \mathbf{V}_{jk} n_{jk}^{-1}) = E(c_2 \chi_{f_2}^2) = c_2 \quad (7)$$

$$V(\text{tr}(\mathbf{Q}_2)) = 2N^2 \sum_{j=1}^a \sum_{k=1}^b \left[\text{vec}(\mathbf{P}_{ll}^*)' \mathbf{V}_{jk}^2 \text{vec}(\mathbf{P}_{ll}^*) \right] \left[n_{jk}^2 (n_{jk} - 1) \right]^{-1} = V(c_2 \chi_{f_2}^2) = 2c_2^2 / f_2.$$

Solving these equations, we have

$$f_2 = \frac{\left[\sum_{j=1}^a \sum_{k=1}^b \text{tr}(\mathbf{P}_{ll}^* \mathbf{V}_{jk} n_{jk}^{-1}) \right]^2}{N^2 \sum_{j=1}^a \sum_{k=1}^b \left[\text{vec}(\mathbf{P}_{ll}^*)' \mathbf{V}_{jk}^2 \text{vec}(\mathbf{P}_{ll}^*) \right] \left[n_{jk}^2 (n_{jk} - 1) \right]^{-1}} \quad (8)$$

If \mathbf{V} is known, f_2 can easily be computed using $N \sum_{j=1}^a \sum_{k=1}^b \text{tr}(\mathbf{P}_{ll}^* \mathbf{V}_{jk} n_{jk}^{-1}) = \text{tr}(\mathbf{P}_t^* \widehat{\mathbf{V}})$ and $N^2 \sum_{j=1}^a \sum_{k=1}^b \left[\text{vec}(\mathbf{P}_{ll}^*)' \mathbf{V}_{jk}^2 \text{vec}(\mathbf{P}_{ll}^*) \right] \left[n_{jk}^2 (n_{jk} - 1) \right]^{-1} = \text{tr}(\widehat{\mathbf{V}} \mathbf{P}_t^{*2} \widehat{\mathbf{D}}^*)$, where $\widehat{\mathbf{D}}^*$ is a diagonal matrix with $1/(n_{jk} - 1)$ as its jk th diagonal element.

In addition to approximating the null distribution of F_N by the central $F(\widehat{f}_1, \widehat{f}_2)$ distribution with estimated df based on Eqs. 6 and 8, respectively (hereafter referred to as BDM), we also focus on an F approximation with estimated numerator df based on Eq. 6 and denominator $df = \infty$ (hereafter referred to as BDM*), as suggested by Brunner et al. (2002) in a nonparametric framework.

Multivariate version of the modified Brown–Forsythe test statistic (MBF)

Practical implementation of the MBF procedure requires estimation of the df of the approximate central p -dimensional Wishart distribution, which can be easily derived by equating the first two moments of the quadratic form associated with m th source of variation in model (1) to those of the central Wishart distribution. A detailed explanation of the so-called multivariate Satterthwaite’s approximation can be found in Vallejo and Ato (2006) in the context of univariate repeated measures. Applying the approach of these authors, the Wilks’s lambda criterion for testing $H_0(t): \mathbf{C}_t \boldsymbol{\mu} = \mathbf{0}$ against $H_0(t): \mathbf{C}_t \boldsymbol{\mu} \neq \mathbf{0}$ can be expressed as the following ratio of two determinants: $\Lambda = |\mathbf{E}_t^*| / |(\mathbf{H}_t + \mathbf{E}_t^*)|$, where the hypothesis matrix (\mathbf{H}_t) and the error matrix (\mathbf{E}_t^*) are determined by

$$\mathbf{H}_t = (\mathbf{C}_t \widehat{\mathbf{B}} \mathbf{A})' \left[\mathbf{C}_t (\mathbf{X}' \mathbf{X})^{-1} \mathbf{C}_t' \right]^{-1} (\mathbf{C}_t \widehat{\mathbf{B}} \mathbf{A}), \quad (9)$$

and

$$\mathbf{E}_t^* = (\widehat{f}_e / \widehat{f}_h) \sum_{j=1}^a \sum_{k=1}^b c_t \widehat{\boldsymbol{\Sigma}}_{jk}^*, \quad (10)$$

where \widehat{f}_e and \widehat{f}_h are the approximate df for \mathbf{E}_t^* and \mathbf{H}_t , respectively, c_t is the jk th diagonal element of the matrix $\mathbf{C}_t' (\mathbf{C}_t \mathbf{D} \mathbf{C}_t')^{-1} \mathbf{C}_t \mathbf{D}$, \mathbf{D} is a diagonal matrix with $1/n_{jk}$ as its jk th diagonal element, and \mathbf{A} is an p -dimensional identity matrix. Using results from Vallejo and Ato, the approximations to the error and hypothesis df , respectively, are

$$\widehat{f}_e = \frac{\left(\text{tr} \sum_{j=1}^a \sum_{k=1}^b c_t \widehat{\boldsymbol{\Sigma}}_{jk} \widehat{\boldsymbol{\Xi}}^{-1} \right)^2 + \text{tr} \left(\sum_{j=1}^a \sum_{k=1}^b c_t \widehat{\boldsymbol{\Sigma}}_{jk} \widehat{\boldsymbol{\Xi}}^{-1} \right)^2}{\sum_{j=1}^a \sum_{k=1}^b \frac{1}{n_{jk}-1} \left[\text{tr}^2 \left(c_t \widehat{\boldsymbol{\Sigma}}_{jk} \widehat{\boldsymbol{\Xi}}^{-1} \right) + \text{tr} \left(c_t \widehat{\boldsymbol{\Sigma}}_{jk} \widehat{\boldsymbol{\Xi}}^{-1} \right)^2 \right]} \quad (11)$$

and

$$\widehat{f}_h = \frac{\left(\text{tr} \sum_{j=1}^a \sum_{k=1}^b c_t \widehat{\boldsymbol{\Sigma}}_{jk} \widehat{\boldsymbol{\Xi}}^{-1} \right)^2 + \text{tr} \left(\sum_{j=1}^a \sum_{k=1}^b c_t \widehat{\boldsymbol{\Sigma}}_{jk} \widehat{\boldsymbol{\Xi}}^{-1} \right)^2}{\sum_{j=1}^a \sum_{k=1}^b M_t + \text{tr}^2 \left(\sum_{j=1}^a \sum_{k=1}^b c_t \widehat{\boldsymbol{\Sigma}}_{jk} \widehat{\boldsymbol{\Xi}}^{-1} \right) + \text{tr} \left(\sum_{j=1}^a \sum_{k=1}^b c_t \widehat{\boldsymbol{\Sigma}}_{jk} \widehat{\boldsymbol{\Xi}}^{-1} \right)^2}, \quad (12)$$

where $M = \left[\text{tr}^2(\widehat{\Sigma}_{jk} \widehat{\Xi}^{-1}) + \text{tr}(\widehat{\Sigma}_{jk} \widehat{\Xi}^{-1})^2 \right] (c_t^\bullet - c_t)$, $\widehat{\Xi} = (c_1^\bullet \widehat{\Sigma}_1 + \dots + c_{ab}^\bullet \widehat{\Sigma}_{ab})$, and $c_t = n_j/N$.

Using the transformation of Wilks’s lambda criterion to F statistic, $H_0(t) : C_t \mu = \mathbf{0}$ ’s are rejected if

$$F_{\text{MBF}_t} = \frac{1 - \widehat{\Lambda}^{1/\widehat{s}}}{\widehat{\Lambda}^{1/\widehat{s}}} \frac{\widehat{f}_2}{\widehat{f}_1} \geq F_{1-\alpha}(\widehat{f}_1, \widehat{f}_2), \tag{13}$$

where $\widehat{f}_1 = t\widehat{f}_h, \widehat{f}_2 = \left[\widehat{f}_e - (t - \widehat{f}_h + 1)/2 \right] s - (t\widehat{f}_h - 2)/2$,

and $\widehat{s} = \left[(t^2 \widehat{f}_h^2 - 4) / (t^2 + \widehat{f}_h^2 - 5) \right]^{\frac{1}{2}}$, with t equal to the dimension of \mathbf{E}_t^* . Here, $F_{1-\alpha}(\widehat{f}_1, \widehat{f}_2)$ is the upper $1-\alpha$ critical value of an F distribution. In addition, we also focus on an F approximation with numerator $df = R(\mathbf{C}_t)$ and estimated denominator df based on Eq. 12 (hereafter referred to as MBF*).

Multivariate version of the general linear model (MLM)

In its classical form, the model defined in Eq. 1 assumes that the different rows of \mathbf{E} are distributed mutually independently and the p elements in any row follow a multivariate normal distribution with mean vector $\mathbf{0}$ and unknown covariance matrix Σ . In other words, while the errors of the l th column of \mathbf{E} are independently distributed as $N(\mathbf{0}, \sigma^2 \mathbf{I}_N)$, the rows of \mathbf{E} have a common covariance matrix Σ that is assumed to be positive definite. Since the previous model is not often very realistic, suppose that we relax the constant variance assumption and assume, instead, that covariance matrices differ from cell to cell. In this case, the matrix \mathbf{V} can be modeled as $\mathbf{V} = \text{diag}\{\mathbf{T}_{11}, \dots, \mathbf{T}_{ab}\}$, where $\mathbf{T}_{jk} = \Sigma_{jk} \mathbf{I}_{n_{jk}}$.

When covariance matrices are unequal across design cells, the appropriate estimator of \mathbf{B} is not the ordinary least squares estimator. For known \mathbf{V} , the appropriate estimator of \mathbf{B} is the generalized least squares (GLS) estimator given by

$$\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \mathbf{X}'\mathbf{V}\mathbf{Y}, \tag{14}$$

with dispersion matrix

$$V(\widehat{\mathbf{B}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \tag{15}$$

If \mathbf{V} is unknown, one may estimate \mathbf{B} , replacing \mathbf{V} with its estimate $\widehat{\mathbf{V}}$ in Eq. 14, which is \mathbf{V} with its \mathbf{T} parameters replaced by their maximum likelihood estimators. Likewise, the variance of $\widehat{\mathbf{B}}$ is usually estimated by replacing \mathbf{V} with its estimate $\widehat{\mathbf{V}}$ in Eq. 15. Although a number of

estimation strategies are available, the present article uses REML estimation as implemented through the SAS Institute’s (2008) PROC MIXED program.

Following Vallejo et al. (2007), the multivariate linear model (hereafter referred to as MLM) described in Eq. 1 can be fitted using PROC MIXED by stacking the p n -vectors \mathbf{y}_i into a single $np \times 1$ vector $\widetilde{\mathbf{y}}$, the p q -vectors β_i into a $pq \times 1$ vector $\widetilde{\beta}$, and the p n -vectors \mathbf{e}_i into a $np \times 1$ vector $\widetilde{\mathbf{e}}$. The resultant vector form can be written as

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{X}}\widetilde{\beta} + \widetilde{\mathbf{e}} \tag{16}$$

where $\widetilde{\mathbf{X}} = \mathbf{I}_p \otimes \mathbf{X}$ is a $np \times pq$ design matrix with $R(\widetilde{\mathbf{X}}) = abp < pq$ for the LFR model and $R(\widetilde{\mathbf{X}}) = abp = pq$ for the FR model. Basically, the multivariate structure is modeled in the PROC MIXED framework by recasting the multivariate model as a univariate one and including an additional level for the multiple outcomes.

An approximated Wald-type F -statistic for testing any null hypothesis of the form $H_0 : \mathbf{R}_t \widetilde{\beta} = \mathbf{0}$ is given by

$$F_t = \frac{1}{f_1} (\mathbf{R}_t \widetilde{\mathbf{b}})' \left[\mathbf{R}_t V(\widetilde{\mathbf{b}}) \mathbf{R}_t' \right]^{-1} (\mathbf{R}_t \widetilde{\mathbf{b}}), \tag{17}$$

where \mathbf{R}_t is a matrix of contrasts of rank f_1 , $\widetilde{\mathbf{b}}$ is the empirical GLS estimator of $\widetilde{\beta}$, and $V(\widetilde{\mathbf{b}})$ is the covariance matrix of $\widetilde{\mathbf{b}}$. It is known that $V(\widetilde{\mathbf{b}})$ tends to underestimate the true sampling variability of $\widetilde{\beta}$ by not accommodating the variability in the estimate of $V(\widetilde{\mathbf{y}})$ when estimating the covariance structure of $\widetilde{\mathbf{b}}$ and when computing associated Wald-type statistics. This is especially important when the sample size is not large enough to support likelihood-based inference and complex covariance structures are used (Kenward, 2001).

To deal with the effect of bias in the estimate of standard errors on inference about the mean structure, Kenward and Roger (1997, 2009) provide a method that can be applied with any fixed effects and covariance structure. This solution involves obtaining an adjusted covariance matrix of fixed effects inflating the conventional estimated covariance matrix of $\widetilde{\mathbf{b}}$, to derive an appropriate F approximation by replacing, in Eq. 12, $V(\widetilde{\mathbf{b}})$ with the adjusted estimator of the covariance matrix of $\widetilde{\mathbf{b}}$, and the estimated denominator df with the generalized F -statistic based on it. The generalized F -statistic approximately follows an F -distribution with numerator df $f_1 = R(\mathbf{R}_t)$ and denominator df f_2 appropriately estimated with the method proposed by Kenward and Roger. A sample syntax for undertaking the MLM analysis is given in the Appendix.

A Monte Carlo study

In order to evaluate the operant characteristics of the proposed approaches, we carried out two simulation studies using a 3×4 factorial MANOVA design. We chose to investigate designs with three and four levels of factors, because these are typical of designs encountered in education and psychology studies. In fact, our preliminary evaluation of the major psychological Spanish journals between January 2001 and December 2008 confirms that the number of levels in a factor typically oscillates between two and six, with a mode of three.

The first study focused on comparing the robustness of the approaches when the normality and covariance homogeneity assumptions did not hold. With this aim the following five variables were manipulated:

1. *Number of dependent variables.* Since some of the procedures examined in this study have been shown to be affected by the number of dependent variables in other contexts (Brunner et al., 2002; Krishnamoorthy & Lu, 2010), we collected Type I error rates when the number of dependent variables was two, three, and four. We initially planned to investigate $p = 6$; however, preliminary simulations indicated that using SAS PROC MIXED took an inordinate amount of time when $p = 6$. It is doubtful that this change substantially affected the results.
2. *Total sample size.* Since the procedures may be affected by the sample size, the performance of the test statistics was investigated using two different global sample size conditions: $N = 108$ and $N = 216$. Only these two sizes were examined, because we felt that they would suffice to provide a comparison between the procedures for small and moderate sample sizes.
3. *Degree of sample size imbalance.* For each sample size condition, both a null and a moderate degree of cell size inequality were explored, as indexed by a coefficient of sample size variation (C), where
$$C = (1/\bar{n}) \left[\sum_{jk} (n_{jk} - \bar{n})^2 / ab \right]^{1/2},$$
 \bar{n} being the average size of the cells. For $N = 108$ and $C = 0$, the size of each cell was (1) $n_{jk} = 9$, and for $N = 108$ and $C = 0.35$, the unequal cell sizes were (2) $n_{1k} = (6, 6, 7, 8)$, $n_{2k} = (6, 8, 10, 11)$, and $n_{3k} = (7, 10, 12, 17)$. When $N = 216$, the sizes of these cells were doubled.
4. *Relationship of cell sizes to dispersion matrices.* Both the positive and negative relationships between cell sizes and dispersion matrices were investigated. In the positive relationship, the larger cell sample sizes were associated with the larger dispersion matrices. In the negative relationship, the smaller cell sample sizes were associated with the larger dispersion matrices. The

heterogeneous covariance matrices we consider here were obtained as follows: $\Sigma_{jk} = (2 \times j \times k) / 2$. Details on these structures when the number of dependent variables was four ($p = 4$) are shown in Table 1.

5. *Population distribution shape.* While the examined procedures are based on the normality assumption, real data rarely conform to normality (Micceri, 1989). Therefore, in order to investigate the possible effects of the shape of the distribution on the robustness of the procedures, we generated data from normal and non-normal distributions, both symmetric and asymmetric, using the methods described in Vallejo, Fernández, Livacic-Rojas, and Tuero-Herrero (2011). Specifically, besides the multivariate normal distribution with univariate skewness (γ_1) and kurtosis (γ_2) equal to zero, the data were obtained from a symmetric light-tailed multivariate distribution with shape parameters equivalent to those of a Laplace distribution (i.e., $\gamma_1 = 0$; $\gamma_2 = 3$), from an asymmetric light-tailed multivariate distribution with shape parameters equivalent to those of an exponential distribution (i.e., $\gamma_1 = 2$; $\gamma_2 = 6$), and from an asymmetric heavy-tailed multivariate distribution with skewness and kurtosis values of 4 and 42, respectively. These values are well within the range of skew and kurtosis encountered in applied psychological research by Micceri.

The second study focused on the comparison of the sensitivity of the approaches for detecting main effects and interactions in two-way MANOVA designs, but only for those methods whose validity does not depend on the underlying assumptions made by them. The data sets were generated in the same way as in the first study, except that they employed different correlation structures between the dependent variables. Because the power of the multivariate test criteria has been shown to be sensitive not only to effect size, but also to the direction and size of the correlations among the dependent variables, we chose to use a small and moderate positive correlation, where all pairwise correlations are .3 and .6, respectively, and a negative correlation where all pairwise correlations between the dependent variables are -.3 or -.6 for $p = 2$ and -.3 for $p = 4$. The different chosen negative correlation values are due to the fact that, for example, a multivariate normal distribution with $p = 4$ and all pairwise correlations of -.6 would not have a positive definite covariance matrix.

When our interest focused on investigating the sensitivity of the tests for detecting the real effect of the rows (variable A), the cell means corresponding to the simple main effect of B at a_1 took on the value 0, and those corresponding to the simple main effect of B at a_2 were

Table 1 Parameter values of the dispersion matrices used to generate the data and types of population mean configurations examined

Unstructured (rho = ± 0.3)	Unstructured (rho = ± 0.6)		
$\Sigma = Cov(e_{ijkl}, e_{ijkl'}) = \sigma_l^2$ if $l = l'$ and $\sigma_{ll'}$ if $l \neq l'$, $\sigma_1^2 = 1; \sigma_2^2 = 2; \sigma_3^2 = 3; \sigma_4^2 = 4;$ $\sigma_{12} = \pm 0.4243; \sigma_{13} = \pm 0.5196; \sigma_{14} = \pm 0.6000;$ $\sigma_{23} = \pm 0.7348; \sigma_{24} = \pm 0.8485; \sigma_{34} = \pm 1.0392;$	$\Sigma = Cov(e_{ijkl}, e_{ijkl'}) = \sigma_l^2$ if $l = l'$ and $\sigma_{ll'}$ if $l \neq l'$, $\sigma_1^2 = 1; \sigma_2^2 = 2; \sigma_3^2 = 3; \sigma_4^2 = 4;$ $\sigma_{12} = \pm 0.8485; \sigma_{13} = \pm 1.0392; \sigma_{14} = \pm 1.2000;$ $\sigma_{23} = \pm 1.4697; \sigma_{24} = \pm 1.6971; \sigma_{34} = \pm 2.0785;$		
Rows $\begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.2 & 0.2 & 0.2 & 0.2 \end{bmatrix}$	$\begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.0 \\ 0.2 & 0.2 & 0.2 & 0.2 \\ 0.4 & 0.4 & 0.4 & 0.4 \end{bmatrix}$	$\begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.0 \\ 0.3 & 0.3 & 0.3 & 0.3 \\ 0.6 & 0.6 & 0.6 & 0.6 \end{bmatrix}$	\cdots $\begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 2.0 & 2.0 & 2.0 & 2.0 \end{bmatrix}$
Columns $\begin{bmatrix} 0.0 & 0.1 & 0.1 & 0.2 \\ 0.0 & 0.1 & 0.1 & 0.2 \\ 0.0 & 0.1 & 0.1 & 0.2 \end{bmatrix}$	$\begin{bmatrix} 0.0 & 0.2 & 0.2 & 0.4 \\ 0.0 & 0.2 & 0.2 & 0.4 \\ 0.0 & 0.2 & 0.2 & 0.4 \end{bmatrix}$	$\begin{bmatrix} 0.0 & 0.3 & 0.3 & 0.6 \\ 0.0 & 0.3 & 0.3 & 0.6 \\ 0.0 & 0.3 & 0.3 & 0.6 \end{bmatrix}$	\cdots $\begin{bmatrix} 0.0 & 1.0 & 1.0 & 2.0 \\ 0.0 & 1.0 & 1.0 & 2.0 \\ 0.0 & 1.0 & 1.0 & 2.0 \end{bmatrix}$
Interaction $\begin{bmatrix} 0.2 & 0.1 & 0.1 & 0.0 \\ 0.0 & 0.1 & 0.1 & 0.2 \\ 0.0 & 0.1 & 0.1 & 0.2 \end{bmatrix}$	$\begin{bmatrix} 0.4 & 0.2 & 0.2 & 0.0 \\ 0.0 & 0.2 & 0.2 & 0.4 \\ 0.0 & 0.2 & 0.2 & 0.4 \end{bmatrix}$	$\begin{bmatrix} 0.6 & 0.3 & 0.3 & 0.0 \\ 0.0 & 0.3 & 0.3 & 0.6 \\ 0.0 & 0.3 & 0.3 & 0.6 \end{bmatrix}$	\cdots $\begin{bmatrix} 2.0 & 1.0 & 1.0 & 0.0 \\ 0.0 & 1.0 & 1.0 & 2.0 \\ 0.0 & 1.0 & 1.0 & 2.0 \end{bmatrix}$

The upper left [2 × 2] matrices were used when the number of dependent variables was two.

based on this value plus the appropriate effect size (i.e., .1, .2, .3, .4, .5, .6, .7, .8, .9 or 1). The remaining cell means were calculated as the mean for factor *B* at *a*₂ plus the appropriate effect size value. On the other hand, when our interest was centered on calculating the power of the columns (variable *B*), the cell means corresponding to the simple main effect of *A* at *b*₁ took on the value 0, and those corresponding to the simple main effects of *A* at *b*₂ and *A* at *b*₃ were based on this value plus the appropriate effect size. The remaining cell means were calculated as the mean for *A* for levels *b*₂ and *b*₃ plus the appropriate effect size value. Finally, when our interest resided in estimating the power of the *AB* interaction, the cell means referring to the intersection of the levels *a*₁*b*₄, *a*₂*b*₁, and *a*₃*b*₁ took on the value 0, and those corresponding to the simple main effects of *A* at *b*₂ and *A* at *b*₃ were based on this value plus the appropriate effect size. The remaining cell means were calculated as the mean for *A* for levels *b*₂ and *b*₃ plus the appropriate effect size value. Details of the procedure described are summarized in Table 1.

Ten thousand replications of each condition were performed using a .05 nominal alpha level. The proportion of false rejections of the null and alternative hypotheses was computed for evaluating the operating characteristics of methods investigated. A SAS/IML macro was used for all calculations.

Results of the robustness study

We used Bradley’s (1978) liberal criterion to facilitate the comparison between our results and those obtained by other researchers in similar studies. According to this criterion, those tests whose empirical error rate ($\hat{\alpha}$) lies in the interval $.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$, would be considered robust. Therefore, for the nominal α level employed in this study ($\alpha = 5\%$), the interval used for defining the robustness of the tests was $2.5 \leq \hat{\alpha} \leq 7.5$. Correspondingly, a test procedure was considered too liberal if its estimated Type I error rate was greater than 7.5 and too conservative if the error rate was smaller than 2.5.

In Tables 2, 3 and 4, we present simulated α -level [in percentages] for the evaluated multivariate tests statistics. Summarizing the simulations, we found the following:

1. When data were obtained from multivariate symmetric distributions, both normal (i.e., $\gamma_1 = 0; \gamma_2 = 0$) and nonnormal (i.e., $\gamma_1 = 0; \gamma_2 = 3$), the descriptive analyses indicated that the error rates for the multivariate versions of the ANOVA-type statistic (i.e., BDM and BDM*), the multivariate versions of the modified Brown–Forsythe procedure (i.e., MBF and MBF*), and the MLM approach were contained within the robustness interval for all the investigated conditions. For $p =$

Table 2 Empirical Type I error rates for the 3×4 factorial MANOVA design ($p = 2$ & $\rho = .6$)

N R (n_{jk} / Σ_{jk})		$(\gamma_1 = 0; \gamma_2 = 0)$			$(\gamma_1 = 0; \gamma_2 = 3)$			$(\gamma_1 = 2; \gamma_2 = 6)$			$(\gamma_1 = 4; \gamma_2 = 42)$		
		H_A	H_B	H_{AB}	H_A	H_B	H_{AB}	H_A	H_B	H_{AB}	H_A	H_B	H_{AB}
MBF													
108	=	4.30	3.82	3.94	3.50	3.12	3.21	3.45	2.43	2.33	3.02	1.86	1.65
216	=	4.47	3.84	3.85	4.02	3.44	3.62	4.18	2.83	2.93	3.49	2.56	2.38
108	+	4.28	3.60	3.65	3.94	3.15	3.24	3.70	2.74	2.49	3.38	2.04	1.74
216	+	4.59	3.93	4.14	4.47	3.42	3.47	4.24	3.14	2.91	4.29	2.46	2.41
108	-	3.53	3.33	3.85	3.03	2.76	2.55	2.89	2.17	2.50	2.67	1.87	1.89
216	-	3.30	3.68	4.17	3.39	2.90	3.37	3.19	2.68	2.62	2.95	2.26	2.48
BMF*													
108	=	5.35	5.49	5.59	5.04	4.81	5.14	5.19	3.90	4.46	4.46	3.68	4.46
216	=	5.48	5.00	6.28	5.29	5.23	5.62	5.84	5.06	5.51	4.98	4.59	4.66
108	+	5.56	5.36	5.72	4.92	4.59	5.24	4.74	4.27	5.01	4.47	3.87	3.80
216	+	5.89	5.18	6.05	5.25	5.27	5.96	5.18	4.91	5.75	4.72	3.93	4.98
208	-	5.05	4.75	5.17	4.04	4.48	4.74	4.51	3.83	3.97	3.83	3.49	3.41
216	-	4.79	4.85	6.01	4.90	4.53	5.77	5.03	4.66	5.46	4.53	4.36	4.66
WJ													
108	=	4.81	5.00	5.38	4.23	4.11	4.26	5.48	5.78	4.38	4.89	4.56	4.77
216	=	4.70	4.83	4.72	4.55	4.16	4.45	5.79	5.53	5.19	4.92	5.12	4.84
108	+	4.73	4.32	5.24	4.06	4.11	4.55	4.59	5.32	4.20	4.21	4.49	4.74
216	+	4.79	4.69	4.51	4.71	4.37	4.21	4.99	5.21	4.53	5.14	4.67	4.37
208	-	4.79	4.83	6.55	4.19	3.93	4.96	5.71	6.06	11.47	5.17	4.93	10.55
216	-	4.70	4.82	5.66	4.40	4.02	4.45	5.62	6.13	9.19	5.12	5.56	8.41
BMD													
108	=	4.90	4.73	4.60	4.56	4.36	4.03	4.20	3.73	2.85	3.83	2.91	2.21
216	=	5.08	4.88	4.97	4.67	4.69	4.33	4.50	4.12	3.78	4.35	3.54	2.78
108	+	4.53	4.82	4.35	4.04	4.17	3.63	4.29	3.41	2.85	3.47	2.97	2.22
216	+	4.75	4.63	4.72	4.69	4.93	4.34	4.76	4.82	3.70	4.17	3.14	2.78
208	-	4.51	4.62	4.45	4.49	4.02	3.81	4.41	3.76	2.76	3.99	3.38	2.48
216	-	4.83	4.75	4.41	4.94	4.56	4.37	4.67	4.11	3.59	4.44	3.63	2.93
BMD*													
108	=	5.42	5.42	5.56	5.10	5.08	5.05	4.87	4.47	3.93	4.95	3.96	3.07
216	=	5.28	5.24	5.31	4.85	4.99	4.71	4.76	4.47	4.13	4.26	3.48	3.12
108	+	4.99	5.38	5.17	4.44	4.85	4.54	4.84	4.06	3.91	4.46	3.68	2.96
216	+	4.91	5.89	5.02	4.87	5.15	4.70	5.00	4.49	4.11	4.32	3.28	3.12
208	-	5.42	4.88	6.13	5.64	5.37	5.43	5.53	5.16	4.31	5.71	4.65	3.78
216	-	5.15	5.21	4.98	5.36	5.11	5.06	5.12	4.62	4.30	4.09	4.45	3.60
MLM													
108	=	4.52	4.54	4.62	4.24	4.02	3.89	4.32	5.36	3.92	3.82	3.96	4.22
216	=	4.74	4.88	5.26	4.74	4.98	3.84	5.74	6.22	4.82	5.28	4.58	4.40
108	+	4.32	4.30	4.50	4.02	4.30	3.00	5.00	4.64	3.76	5.36	4.06	3.74
216	+	4.64	4.78	4.76	4.08	4.03	3.91	5.68	5.57	4.60	3.92	4.24	4.58
108	-	4.50	4.00	4.62	4.30	3.38	3.10	6.66	5.60	8.88	5.21	4.80	7.50
216	-	5.54	4.70	4.48	4.90	4.21	3.67	5.94	5.12	8.58	5.20	4.92	8.48
PROC GLM													
108	=	5.90	5.87	6.74	5.56	5.40	6.46	5.76	5.65	6.00	5.58	4.79	5.58
216	=	5.85	5.60	6.12	5.55	5.26	6.33	5.87	5.11	6.27	5.51	5.16	5.91
108	+	1.61	1.18	1.40	1.35	1.23	1.37	1.86	1.36	1.35	1.85	1.19	1.33
216	+	1.67	1.19	1.41	1.73	1.05	1.20	1.78	1.27	1.26	2.10	1.36	1.43

Table 2 (continued)

N	R (n_{jk} / Σ_{jk})	$(\gamma_1 = 0; \gamma_2 = 0)$			$(\gamma_1 = 0; \gamma_2 = 3)$			$(\gamma_1 = 2; \gamma_2 = 6)$			$(\gamma_1 = 4; \gamma_2 = 42)$		
		H _A	H _B	H _{AB}	H _A	H _B	H _{AB}	H _A	H _B	H _{AB}	H _A	H _B	H _{AB}
108	–	11.98	14.14	16.90	12.01	13.85	15.85	11.68	13.62	15.79	10.52	12.19	14.73
216	–	11.41	14.15	16.49	11.26	13.18	15.81	11.63	13.32	15.26	10.85	12.27	14.23

N, total sample size; $R(n_{jk} / \Sigma_{jk})$, relationship between the size of the cells and the size of the covariance matrices. Bold values are outside the [2.50–7.50] interval.

2, the WJ procedure also maintains rates of Type I error closer to the nominal value. However, for $p = 3$ and $p = 4$, it tended to have an inflated Type I error rate that improved as the sample size increased. On the other hand, it was also found that the MANOVA test using the transformation of Wilks's lambda criterion to F statistic shows robustness properties similar to those of the ANOVA test, although the higher the dimensionality the less robust.

- When data were obtained from an asymmetric light-tailed multivariate distribution (i.e., $\gamma_1 = 2; \gamma_2 = 6$), the descriptive analyses indicated that the error rates for the BDM, BDM*, and MBF* procedures were contained within the robustness interval for all the investigated conditions. The rates of Type I error for the MBF procedure were frequently conservative. The tendency to be conservative was worse when sample sizes were small. On the other hand, the WJ and MLM procedures were occasionally liberal for the test of the interaction effect when the smaller cell sample sizes were associated with the larger dispersion matrices, with the degree of liberalness decreasing with increases in the sample sizes. With respect to the MANOVA test, the results shown in the tables indicate a substantial effect of the presence of covariance heterogeneity and a very slight effect of the absence of multivariate normality.
- When data were obtained from an asymmetric heavy-tailed multivariate distribution (i.e., $\gamma_1 = 4; \gamma_2 = 42$), the descriptive analyses indicated that the error rates for the BDM* and MBF* procedures were contained within the robustness interval for all the investigated conditions. The rates of Type I error for the BMD and MBF procedures were frequently very conservative, while the rates of the WJ and MLM procedures were occasionally liberal for the test of the interaction effect. Also, not surprisingly, the pattern of results obtained with the MANOVA test was also not affected by the shape of the distribution. Once more, the critical variable was the absence of covariance homogeneity.

Results of the study of power

When data were obtained from asymmetric multivariate distributions, both light-tailed (i.e., $\gamma_1 = 2; \gamma_2 = 6$) and heavy-tailed (i.e., $\gamma_1 = 4; \gamma_2 = 42$), error rates of the BDM* and MBF* procedures were scarcely affected by the lack of homogeneity. It is also clear from Tables 2, 3 and 4 that the MLM approach controls Type I error rates in most of the investigated conditions. The other methods could not be recommended because of their conservatism or liberalism. Therefore, we focused on comparing the BDM*, MBF*, and MLM approaches with regard to their sensitivity in detecting differences between nonnull mean vectors of the two-way MANOVA design.

To assess the power of the BDM*, MBF*, and MLM tests, we manipulated the number of dependent variables ($p = 2$ or $p = 4$), total sample size ($N = 108$ or $N = 216$), degree of sample size imbalance ($C = 0$ or $C = .35$), relationship of cell sizes to dispersion matrices (positive or negative), effect sizes (.1, .2, . . . or 1), direction and size of the correlations among the dependent variables ($\pm .3$ or $\pm .6$), and population distribution shape (normal or exponential). Preliminary simulations suggested that little would be learned beyond the conclusions reached using the normal and exponential distributions, so the performance for the Laplace distribution was not examined. The power of the three analytical procedures was estimated on each effect size for unweighted row and column effects and for testing interaction effects. For tests of the interaction, the shape of these curves, averaging across the total sample size, degree of sample size imbalance, and relationship of cell sizes to dispersion matrices, are displayed graphically (Figs. 1 and 2).

The power results of the Monte Carlo experiments displayed in Figs. 1 and 2, can be summarized as follows:

- Overall, the analytical procedures performed better when the total sample size and/or the effect sizes increased. The procedures also performed generally better when group sizes and covariance matrices were positively paired than when they were negatively paired. Averaging across all the manipulated condi-

Table 3 Empirical Type I error rates for the 3×4 factorial MANOVA design ($p = 3$ & $\rho = .6$)

N R (n_{jk} / Σ_{jk})		$(\gamma_1 = 0; \gamma_2 = 0)$			$(\gamma_1 = 0; \gamma_2 = 3)$			$(\gamma_1 = 2; \gamma_2 = 6)$			$(\gamma_1 = 4; \gamma_2 = 42)$		
		H_A	H_B	H_{AB}	H_A	H_B	H_{AB}	H_A	H_B	H_{AB}	H_A	H_B	H_{AB}
MBF													
108	=	4.16	3.69	4.28	3.64	3.01	3.11	3.53	2.54	2.16	2.85	2.12	1.77
216	=	4.17	3.73	3.89	4.17	3.45	3.72	4.21	3.22	2.09	3.39	2.59	2.38
108	+	4.90	3.75	3.71	4.08	2.90	3.12	3.80	2.46	1.96	3.07	2.13	1.66
216	+	4.60	3.97	3.95	4.42	3.91	3.89	4.30	3.43	3.31	3.69	2.50	2.22
108	-	3.52	3.11	3.31	3.11	2.56	2.97	2.65	2.40	2.06	2.58	2.07	1.71
216	-	4.18	3.50	4.03	3.41	3.42	3.39	3.33	2.73	2.99	2.69	2.27	2.13
BMF*													
108	=	5.48	5.21	5.26	4.99	4.69	5.30	5.20	4.77	4.20	4.24	4.00	3.53
216	=	5.61	5.48	6.39	5.55	5.15	5.00	5.58	5.50	5.32	4.99	4.49	4.55
108	+	5.62	5.05	5.51	5.52	4.79	5.01	4.89	4.20	5.69	4.94	3.36	4.18
216	+	5.28	5.40	6.12	5.58	5.06	5.99	5.31	4.73	5.09	4.83	4.57	5.04
208	-	4.90	4.84	5.01	3.98	4.76	4.51	4.12	3.74	3.67	3.84	3.38	3.23
216	-	5.22	5.18	5.94	4.59	4.00	5.22	5.15	4.98	5.05	4.00	4.06	4.26
WJ													
108	=	4.69	4.97	7.02	4.23	3.90	4.27	5.70	5.85	5.13	4.53	5.18	5.64
216	=	4.68	4.71	5.29	4.86	4.12	4.22	5.31	6.05	4.77	5.17	5.28	4.39
108	+	4.86	4.95	6.23	4.14	3.89	4.85	4.62	4.43	4.61	3.94	4.38	5.13
216	+	4.80	4.92	5.03	4.39	4.50	4.45	4.89	5.20	5.16	5.17	4.36	4.79
208	-	4.77	4.02	9.60	4.01	4.28	7.56	5.73	5.96	15.60	5.03	5.59	14.37
216	-	4.99	4.54	5.79	4.40	4.71	5.08	6.22	6.03	11.02	5.17	5.31	8.88
BMD													
108	=	5.08	4.81	4.39	4.35	4.13	3.80	4.50	3.62	2.75	3.76	3.65	2.18
216	=	4.89	5.03	4.84	4.67	4.75	4.33	4.58	4.18	3.64	4.37	3.60	2.62
108	+	4.81	4.49	4.23	4.57	4.27	3.61	4.27	3.65	2.74	3.51	2.84	1.98
216	+	4.63	5.04	4.60	4.76	4.53	4.32	4.56	4.26	3.90	3.73	3.23	2.58
208	-	4.72	4.29	4.15	4.16	3.81	3.59	4.61	3.38	2.41	3.98	3.25	2.31
216	-	4.67	4.71	4.91	4.68	4.32	4.18	4.68	4.18	3.32	4.27	3.58	2.80
BMD*													
108	=	5.61	5.33	5.37	4.87	4.77	4.76	5.03	4.40	3.69	4.44	3.63	2.95
216	=	5.07	5.28	5.97	4.84	5.00	4.64	4.81	4.53	4.14	4.62	3.86	2.96
108	+	5.18	5.12	5.42	5.07	4.93	4.50	4.77	4.27	3.60	4.01	3.52	2.67
216	+	4.84	5.29	4.27	4.98	4.73	4.63	4.77	4.56	4.27	3.96	3.51	2.98
208	-	5.59	5.28	5.25	5.07	5.04	4.96	5.80	4.64	3.79	4.90	4.38	3.48
216	-	5.00	5.05	5.74	5.16	4.82	4.87	5.01	4.72	4.03	4.57	4.08	3.32
MLM													
108	=	4.08	4.73	5.05	3.92	3.70	3.28	4.74	5.04	2.58	4.34	4.70	2.64
216	=	4.66	5.22	5.24	4.82	4.82	4.04	5.68	6.00	4.68	5.28	4.90	3.78
108	+	4.38	4.32	4.20	4.08	3.30	2.82	5.26	4.86	2.52	3.90	3.58	2.98
216	+	5.56	5.48	5.04	4.64	4.44	4.06	4.88	5.12	4.96	4.44	4.06	3.92
108	-	3.69	4.58	3.97	3.86	3.56	3.72	5.44	4.25	8.99	4.50	4.32	7.96
216	-	5.46	4.67	5.30	4.36	4.24	4.06	6.26	6.40	9.63	5.48	4.58	8.22
PROC GLM													
108	=	5.89	6.04	7.10	5.91	5.60	6.68	6.05	5.54	6.22	5.39	5.18	5.77
216	=	5.83	5.77	6.64	5.93	5.46	6.58	5.95	5.72	6.31	5.72	5.12	5.70
108	+	1.48	0.89	1.03	1.13	0.88	0.85	1.67	0.80	1.00	1.57	1.12	1.03
216	+	1.41	0.85	0.79	1.12	1.07	0.94	1.13	1.12	1.07	1.66	0.93	1.05

Table 3 (continued)

N R (n_{jk} / Σ_{jk})	$(\gamma_1 = 0; \gamma_2 = 0)$			$(\gamma_1 = 0; \gamma_2 = 3)$			$(\gamma_1 = 2; \gamma_2 = 6)$			$(\gamma_1 = 4; \gamma_2 = 42)$			
	H _A	H _B	H _{AB}	H _A	H _B	H _{AB}	H _A	H _B	H _{AB}	H _A	H _B	H _{AB}	
108	–	13.86	16.92	19.13	13.31	16.19	19.20	13.34	15.93	18.06	12.89	14.68	17.74
216	–	13.32	16.01	19.08	13.39	16.43	19.15	13.40	15.57	18.18	12.05	15.09	17.74

Note. See note from Table 2.

- tions, the overall success was 22.4% for the BDM*, 36.3% for the MBF*, and 45.0% for the MLM.
- For data with positively correlated variables, the differences between the MBF* and BDM* were less than 1 percentage point regardless of number of dependent variables, population distribution shape, or size of the correlations among the dependent variables investigated. With positive correlation, MLM turned out to be uniformly superior to its competitors, correctly rejecting the null hypothesis 29.3% of the time. The percentage of rejections achieved by the rest of the approaches was 20.7% with BDM* and 21.4% with MBF*.
 - For data with negatively correlated variables, the MBF* and MLM procedures performed much better than the BDM* approach, which was generally little affected by the direction of the correlations among the dependent variables. With negative correlation, MLM turned out to be uniformly superior to its competitors, correctly rejecting the null hypothesis 60.7% of the time. The percentage of rejections achieved by the rest of the approaches was 24.1% with BDM* and 51.2% with MBF*.
 - When the different variables were positively correlated, the results show that the power of all three tests declined slightly as the number of dependent variables increased, with differences ranging from 4.5 to 6.7 percentage points. In the case of positively correlated variables, we also found that as the correlation between dependent variables increased, the power of the tests decreased slightly, with differences ranging from 0.7 to 2.8 percentage points. In contrast, when the different variables were negatively correlated, the power of the MBF* and MLM tests increased moderately as the number of dependent variables increased, with differences ranging from 13.2 to 14.5 percentage points. In addition, stronger relationships among the variables were associated with higher power, particularly for smaller numbers of dependent variables, with differences ranging from 14.0 to 17.2 percentage points.
 - Finally, the simulated power curves displayed in Figs. 1 and 2 reveal that the pattern of power for the MLM

approach was slightly affected by the form of distribution. Specifically, when data were generated from an asymmetric distribution with a moderate degree of kurtosis (i.e., exponential-type data), the MLM approach was slightly more powerful than the other distribution (i.e., multivariate normal), although the averaged power differences did not reach 5 percentage points. However, it can be seen that the power rates of the BDM* and MBF* tests scarcely affected the form of distribution.

Basically, the pattern of results found when our interest was centered on comparing the sensitivity of the methods for detecting the row and column effects was qualitatively similar to the pattern shown in Figs. 1 and 2 for the interaction. Hence, in order to conserve space, the results are neither presented nor discussed here but are available from the first author.

Discussion and recommendations

The operating characteristics of the three approximate *df* solutions based on ordinary least squares (i.e., MBF, BDM, and WJ) were compared with an empirical form of generalized least squares method (i.e., MLM) under violations of normality and covariance homogeneity in a multivariate factorial design setup. The BDM and WJ approaches used in this article constitute a direct extension of the works of Brunner et al. (1997) and Johansen (1980), respectively. In the specific context of univariate factorial designs, previous studies had revealed that the two procedures were generally robust to violations of underlying assumptions (Keselman, Carriere, & Lix, 1995; Vallejo et al., 2010a, b). On the other hand, MBF and MLM have also been proven to be robust when they have been used to analyze longitudinal data with unequal dispersion matrices (Vallejo & Ato, 2006; Vallejo et al., 2007). Our results are consistent with those obtained in the aforementioned studies; furthermore, they provide new findings that help researchers in the selection of feasible alternatives for testing main and interaction effects.

Table 4 Empirical Type I error rates for the 3×4 factorial MANOVA design ($p = 4$ & $\rho = .6$)

N R (n_{jk} / Σ_{jk})		$(\gamma_1 = 0; \gamma_2 = 0)$			$(\gamma_1 = 0; \gamma_2 = 3)$			$(\gamma_1 = 2; \gamma_2 = 6)$			$(\gamma_1 = 4; \gamma_2 = 42)$		
		H_A	H_B	H_{AB}	H_A	H_B	H_{AB}	H_A	H_B	H_{AB}	H_A	H_B	H_{AB}
MBF													
108	=	4.72	3.74	3.53	3.42	2.79	2.91	3.72	2.69	2.08	2.88	1.89	1.68
216	=	4.70	3.92	4.21	4.55	3.46	3.62	4.52	3.23	2.47	3.53	2.10	2.13
108	+	4.45	3.66	3.55	3.04	3.12	3.05	3.56	2.46	2.03	3.27	2.07	1.72
216	+	4.98	3.94	3.85	4.44	3.89	3.90	4.36	3.20	3.17	3.85	2.53	1.94
108	-	3.53	3.25	3.56	3.15	2.47	2.85	2.40	2.04	2.05	2.50	1.93	1.86
216	-	3.95	3.50	3.83	3.60	3.22	3.41	3.36	2.70	3.11	2.88	2.38	2.12
BMF*													
108	=	5.23	5.21	5.50	4.60	4.57	4.53	4.98	4.35	4.23	4.02	3.40	4.43
216	=	5.32	5.15	6.13	5.80	5.29	5.63	5.93	4.76	5.17	4.95	4.29	4.75
108	+	5.18	5.08	5.41	5.01	4.78	5.52	4.43	4.40	4.45	3.97	3.31	3.44
216	+	5.88	5.11	6.34	5.70	5.77	5.71	5.65	4.84	5.22	4.71	4.20	4.59
208	-	4.67	4.35	5.04	4.33	3.73	4.29	3.89	3.36	3.79	3.45	3.21	2.89
216	-	5.14	5.00	5.55	4.85	4.73	4.62	5.17	4.63	4.90	4.19	3.61	4.34
WJ													
108	=	5.02	5.28	8.51	3.85	4.55	6.79	5.47	5.79	6.24	4.50	5.19	7.13
216	=	5.01	4.75	5.54	4.59	4.65	4.92	5.92	5.82	5.60	5.11	4.46	4.98
108	+	4.57	5.10	7.93	4.35	4.10	6.77	4.40	4.83	6.79	3.98	4.40	6.85
216	+	4.89	4.97	5.55	4.60	4.54	4.78	4.95	5.40	5.14	4.50	4.66	4.09
208	-	4.81	6.34	15.93	3.02	5.12	13.06	5.94	6.70	23.25	5.46	6.01	19.95
216	-	5.07	5.77	7.37	4.45	4.38	5.85	6.22	5.91	13.59	5.40	5.43	10.34
BMD													
108	=	4.67	4.59	4.41	4.65	4.41	3.76	4.28	3.58	2.69	3.74	2.80	2.09
216	=	5.03	5.04	4.90	4.61	4.53	4.46	4.84	4.52	3.71	4.39	3.28	2.64
108	+	4.60	4.59	4.00	4.58	4.29	3.67	4.11	3.31	2.89	3.28	2.63	1.84
216	+	4.86	4.95	4.79	4.51	4.69	4.35	4.57	4.17	3.67	4.19	3.25	2.44
208	-	4.71	4.50	4.00	4.20	3.82	3.31	4.63	3.33	2.52	3.92	2.92	2.02
216	-	4.70	4.78	4.65	4.68	4.26	3.98	4.66	3.11	3.41	4.35	3.59	2.75
BMD*													
108	=	5.25	5.00	5.27	5.04	5.03	4.70	4.91	4.27	3.60	4.40	3.39	2.81
216	=	5.14	5.29	5.17	4.76	4.82	4.69	4.99	4.78	4.06	4.57	3.54	2.94
108	+	5.01	5.08	4.82	5.04	4.89	4.55	4.63	3.96	4.73	3.88	3.35	2.52
216	+	5.00	5.11	5.04	4.65	4.87	4.63	4.71	4.42	4.00	4.44	3.60	2.75
208	-	5.31	5.55	5.31	5.14	4.92	4.84	5.08	4.47	3.81	4.78	4.00	3.10
216	-	4.97	5.27	5.28	5.04	4.71	4.48	4.95	4.59	3.99	4.69	3.59	3.28
MLM													
108	=	4.50	3.76	4.49	3.80	3.84	3.08	5.08	4.56	2.53	4.08	4.56	2.62
216	=	4.70	4.58	5.52	4.08	4.40	3.80	5.76	5.33	4.57	4.76	4.40	3.76
108	+	4.24	4.10	4.43	4.88	2.92	2.75	4.44	3.08	3.16	4.32	3.84	2.77
216	+	5.25	5.10	4.81	4.96	3.96	3.84	5.16	5.20	5.00	4.40	4.64	3.40
108	-	3.57	4.74	4.12	3.28	2.88	3.96	6.04	4.64	9.08	4.08	6.88	7.69
216	-	4.70	4.59	5.13	4.20	4.68	3.80	6.52	5.40	9.70	5.00	4.52	8.01
PROC GLM													
108	=	6.84	6.16	6.93	5.66	5.60	6.36	6.56	5.71	5.94	5.90	5.16	5.78
216	=	6.10	5.99	6.80	6.37	5.90	6.36	6.70	5.66	6.72	5.80	4.65	5.50
108	+	1.06	0.80	0.72	1.15	0.80	0.63	1.42	0.76	0.73	1.25	0.81	0.68
216	+	1.18	0.63	0.91	1.18	0.68	0.78	1.22	0.81	0.70	1.04	0.84	0.56

Table 4 (continued)

N R (n_{jk} / Σ_{jk})	$(\gamma_1 = 0; \gamma_2 = 0)$			$(\gamma_1 = 0; \gamma_2 = 3)$			$(\gamma_1 = 2; \gamma_2 = 6)$			$(\gamma_1 = 4; \gamma_2 = 42)$			
	H _A	H _B	H _{AB}	H _A	H _B	H _{AB}	H _A	H _B	H _{AB}	H _A	H _B	H _{AB}	
108	–	15.33	19.88	23.37	15.33	19.87	22.24	15.50	17.87	21.02	14.16	16.22	19.19
216	–	15.92	18.78	22.76	14.92	17.28	21.50	14.31	17.28	20.58	13.95	16.62	19.13

See note from Table 2.

With respect to robustness, it is noteworthy that of the evaluated methods, only the BDM* and MBF* procedures performed satisfactorily under every condition manipulated. When the data were extracted from multivariate symmetric distributions, the BDM, BDM*, MBF,

MBF*, and MLM approaches were much closer to the nominal alpha level over the variety of conditions considered in the experiments than were the WJ and PROC GLM tests. Rates of Type I error for the PROC GLM were very conservative (below 0.6%) or very liberal

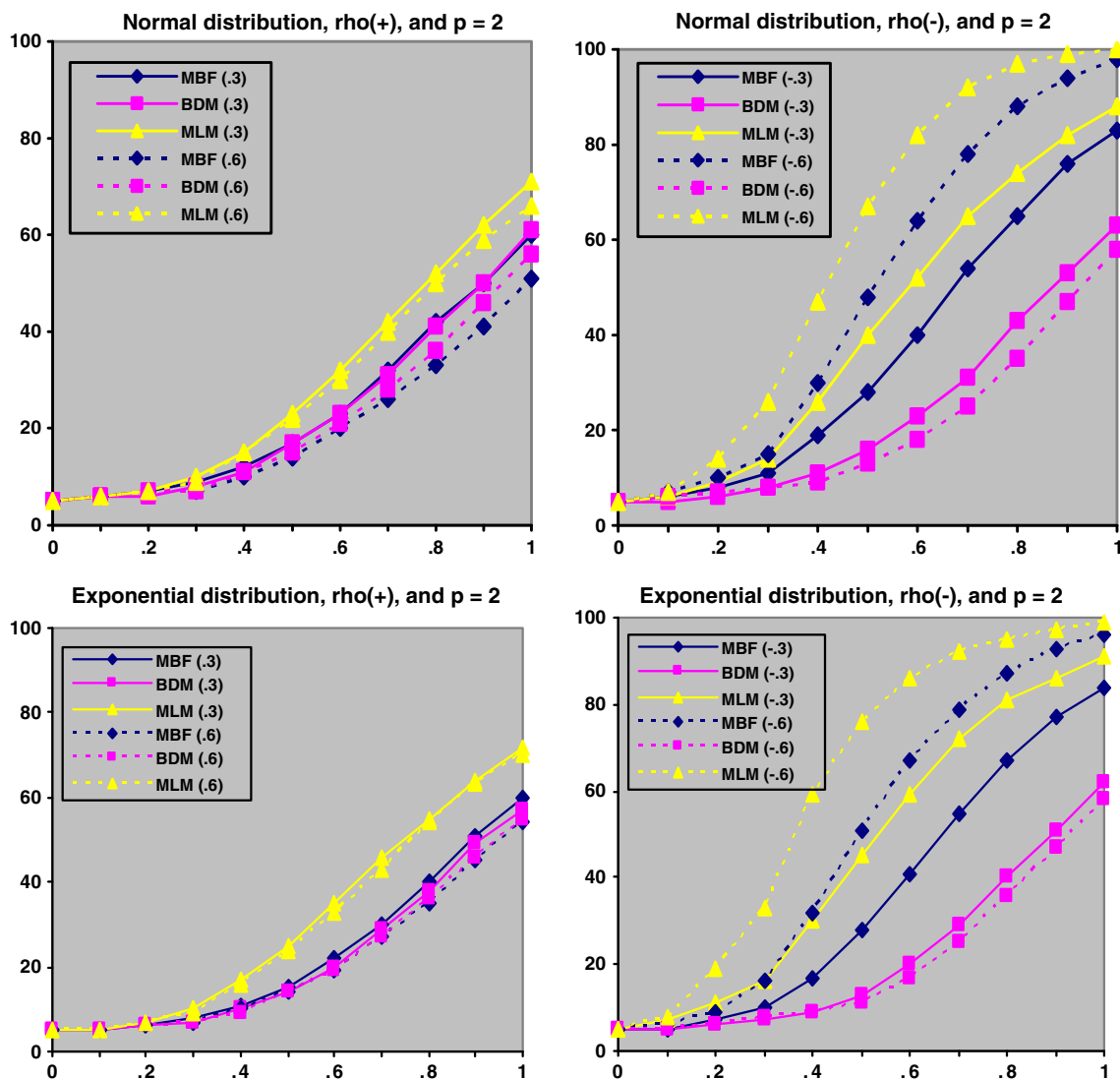


Fig. 1 Estimated power curves corresponding to the interaction ($p = 2$), averaging across the total sample sizes and type of pairing of cell sizes, for the robust multivariate tests in comparison: MBF*, BDM*,

and MLM. The underlying distribution is multivariate normal (above) or multivariate exponential (below) with positive correlation (left) or negative correlation (right)

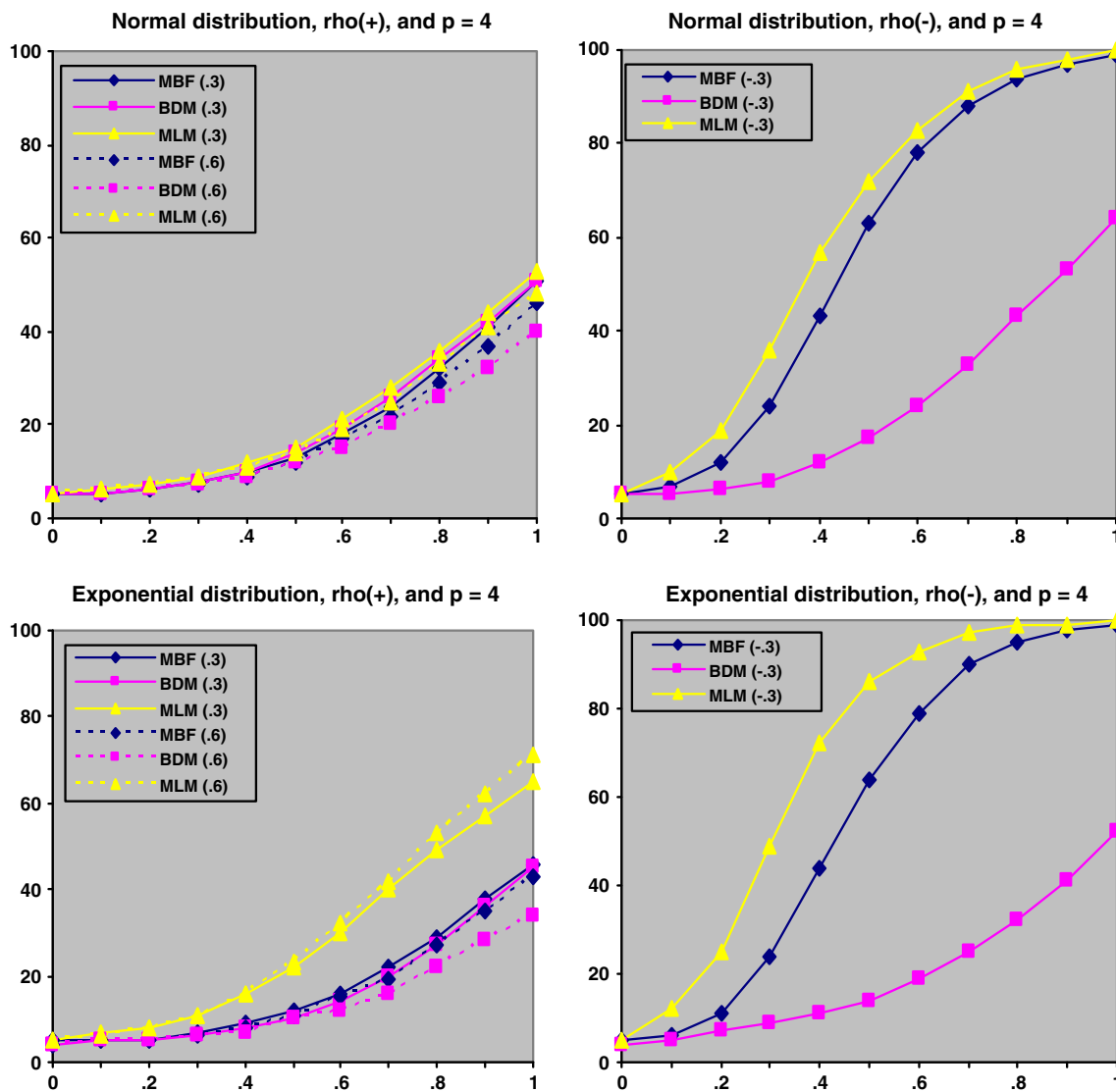


Fig. 2 Estimated power curves corresponding to the interaction ($p = 4$), averaging across the total sample sizes and type of pairing of cell sizes, for the robust multivariate tests in comparison: MBF*, BDM*,

and MLM. The underlying distribution is multivariate normal (above) or multivariate exponential (below) with positive correlation (left) or negative correlation (right)

(above 23%) when cell sizes and covariance matrices were positively or negatively paired, respectively, while the performance of the WJ procedure for tests of the interaction was clearly affected by the type of pairing of cell sizes and covariance matrices, as well as the number of dependent variables and the total sample size. However, for the null hypothesis of zero row and column effects, the WJ procedure provided adequate control of the Type I error rates for all the situations considered.

When data were obtained from multivariate asymmetric distributions, our results also indicated that both BDM* and MBF* procedures were able to provide very effective Type I error control, while the rates for the BDM and MBF approaches were often conservative, especially when the data were generated from extremely biased distributions. It

is also important to note that the executions PROC GLM were again markedly affected by the lack of homogeneity among the covariance matrices and scarcely by the lack of normality, while the remaining procedures studied (i.e., MLM and WJ) tended to yield liberal values for tests of the interaction under negatively paired conditions. On the whole, the results revealed that the MLM approach provided better Type I error control than did the WJ test, but it also becomes liberal. Although previous findings (Krishnamoorthy & Lu, 2010) indicated that rates of error for the WJ were frequently well controlled for the one-way MANOVA, the data from this study demonstrate that the WJ procedure was often liberal if the interactive effects were involved in the design, particularly when the number of dependent variables increased and total sample size was

small, an observation made by Lix et al. (2003) in the context of multivariate repeated measures data.

With respect to power, the results of our study for tests of the interaction, commonly the most interesting issue for researchers (Maas & Snijders, 2003), showed that the MLM procedure was uniformly more powerful than its most direct competitors. In fact, averaging across all the conditions manipulated, the overall success was 22.4% for the BDM*, 36.3% for the MBF*, and 45.0% for the MLM. The BDM* and MBF* procedures behaved similarly for data with positively correlated variables regardless of the number of dependent variables, the population distribution shape, or the size of the correlations among the dependent variables investigated. However, the BDM* procedure always gave less power than did the MBF* procedure for data with negatively correlated variables. Of course, these power differences are not at the expense of high Type I error rates, since both procedures provide similar control of the error rates. Hence, on the basis of our results and those reported by Bathke, Harrar, and Madden (2008), we recommend using the modified-type ANOVA statistic with caution.

Given that the MLM approach was only occasionally liberal when cell sizes and covariance matrices were negatively paired, and then by only a small amount, and that the MLM was uniformly more powerful than the BDM* and MBF* tests, the MLM would be recommended for most applied researchers. Only if it were crucial to maintain strict alpha control would a researcher choose BDM* or MBF* and then MLM. Consequently, on the basis of the power analysis, it is reasonable to consider that the MLM approach constitutes an appropriate solution for analyzing two-way MANOVA models when the underlying assumptions are not met. In addition, it should not be forgotten that this approach can easily be performed with standard statistical packages, such as SAS (SAS Institute, 2008) PROC MIXED.

Notwithstanding the above, one reviewer noted that the degree of covariance heterogeneity used in this article seems small. As was mentioned above, the largest and smallest covariance matrices differed by a 12:1 ratio. This ratio may seem considerable; however, Keselman et al. (1998) noted that in a review of articles published in prominent education and psychology journals, ratios as large as 24:1 and 29:1 were observed in factorial completely randomized designs. In response, we reanalyzed the robustness of BDM*, MBF*, and MLM methods by varying the degree of covariance heterogeneity. To have some reassurance that these methods will perform well for data encountered in practice, we examined their performance when the unequal covariance matrices were in the

ratios 16:1, 25:1, and 36:1, and the data were obtained from the multivariate normal and multivariate nonnormal distributions considered in this article. As was expected, the results obtained using the heteroscedastic methods based on ordinary least squares (i.e., MBF, BDM) were very similar to those reported in Tables 2, 3 and 4, so the details are not reported here but are available from the authors upon request. Perhaps the most important result found in this reanalysis was that under normality, the empirical form of generalized least squares method (i.e., MLM) might be a reasonable strategy. If moderate to severe skewness is present, however, (e.g., $\gamma_1 = 2$; $\gamma_2 = 6$ or $\gamma_1 = 4$; $\gamma_2 = 42$), the MLM method can, on occasion, be nonrobust.

To determine the generalizability of our findings, we also revised the robustness of the BDM*, MBF*, and MLM methods by varying the degree of kurtosis (i.e., $\gamma_1 = 0$ and $\gamma_2 = 12, 24, 36,$ and 54). The results reached using three dependent variables ($p = 3$) and a ratio of (largest to smallest) standard deviations of 5 (these results are available from the authors upon request) show that Type I error rates for the symmetric moderate-tailed distributions ($\gamma_1 = 0$ and $\gamma_2 = 12$ and 24) are reasonably close to 0.05, but it is also found that the methods become conservative as the distributions become more heavy-tailed. In fact, for symmetric heavy-tailed distributions ($\gamma_1 = 0$ and $\gamma_2 = 36$ and 54), the three methods tended to produce conservative results. Without entering into a more extended discussion, it is important for the interpretation of results to note that the MBF* method was more conservative for negative than for positive pairing conditions, while the BDM* and MLM methods were frequently conservative for several of the conditions of positive and negative pairings of group sizes and covariance matrices, and the interaction error rates were almost always below those for the MBF* method. It is also important to note that with kurtosis values equal to or greater than 36, the sample size required by these methods to achieve robustness could be much larger than researchers are likely to have available to them in the behavioral sciences.

To conclude, we would like to add four comments. First, the pattern of results was very clear and consistent, so we feel that they could be generalized to a wider range of conditions. Nevertheless, readers should keep in mind that our findings are suggestive as to what might happen in several heteroscedastic and nonnormal cases, but not definitive for all types of heterogeneity and nonnormality. Second, if the distribution from which the data are sampled is heavy-tailed, the adverse effects of nonnormality can probably be overcome by substituting robust measures of location (e.g., trimmed mean) and scale (e.g., Winsorized covariance matrices) for the usual mean and covariance matrices. According to Keselman, Algina, Lix, Wilcox, and

Deering (2008), one argument for the trimmed mean is that it can have a substantial advantage in terms of accuracy of estimation when sampling from heavy-tailed symmetric distributions without altering the hypothesis tested, because it represents the center of the data. In particular, Wilcox (2005) has shown that with modest sample sizes, the 20% trimmed mean performs well in many situations because it is able to handle a high proportion of outliers. Third, if the distribution from which the data are sampled is skewed, choosing a completely valid alternative is complex. One possible solution could be to use a predetermined amount of trimming and symmetrizing transformations (see Luh & Guo, 2004; Vallejo et al., 2010b) to contend with skewness and heteroscedastic statistics to contend with heterogeneity (e.g., the modified versions of the MBF and BDM methods). It is also possible to use asymmetric trimming, determining the amount of trimming in each tail on the basis of descriptive measures of distribution shape. Another option for overcoming the deleterious effects of nonnormality is to use permutation and bootstrapping methods, not forgetting the solution that involves replacing the original data with ranks and then applying the heteroscedastic techniques considered in this article. Lastly, in future research, it would also be informative to examine the performance of the linear model using techniques that allow distribution of error terms other than the normal ones and relax the requirement of constant variability. The SAS PROC GLIMMIX module, a generalization of the MIXED and GENMOD procedures, allows researchers to fit a sequence of models to data when normality and covariance assumptions are not necessarily satisfied (see Schabenberger, 2007, for details and suggestions). Nonparametric resampling methods, such as the bootstrap and permutation, as well as the MLM approach conducted through SAS PROC GLIMMIX have not been evaluated in comparative studies, but they may be able to offer competitive results.

Author Note We gratefully thank the associate editor and the anonymous reviewers for many constructive comments that led to substantial improvements in the manuscript. This work was supported by grants from the Spanish Ministry of Science and Innovation (Ref.: PSI-2011-23095) and the Spanish Ministry of Education and Science (Ref.: SEJ-2006-09025).

Appendix: Sample SAS commands

To illustrate the analysis of the heteroscedastic two-way MANOVA model, as implemented in the SAS® (2008,

version 9.2) PROC MIXED, we use the hypothetical data of 24 subjects. The experiment involved a 3×4 design with two replications, for a total of 24 observation vectors.

SAMPLE DATA

ID	A	B	Y1	Y2
1	1	1	46	32
2	1	1	26	82
3	1	2	22	44
4	1	2	16	32
5	1	3	26	52
6	1	3	96	72
7	1	4	32	64
8	1	4	26	52
:	:	:	:	:
17	3	1	26	52
18	3	1	36	72
19	3	2	32	64
20	3	2	46	52
21	3	3	36	72
22	3	3	46	92
23	3	4	42	34
24	3	4	56	72

Where **ID** denotes the specific subject, **A** and **B** indicate row and column variable levels, respectively, and **Y1–Y2** represents the multivariate observations. The following SAS code could be used for changing multivariate data format to MIXED data format,

```
DATA MANOVA; ID= _N_;
ARRAY V{2} Y1 Y2;
K=1;
DO VAR = 1 TO 2;
Y = V {K};
OUTPUT;
K = K + 1;
END;
KEEP ID A B K VAR Y;
RUN;
```

where **K** and **VAR** are extra variables levels. The multivariate analysis was carried out using SAS PROC MIXED with the following statements:

```
PROC MIXED DATA=MANOVA METHOD=REML;
CLASS ID A B K VAR;
MODEL Y=VAR VAR*A|VAR*B/NOINT DDFM=KR HTYPE=3;
REPEATED K/TYPE=UN GROUP=A*B SUB=ID(A B);
ODS OUTPUT TESTS3=PROBF;
ODS LISTING CLOSE;
RUN;
ODS LISTING;
```

References

- Algina, J. (1994). Some alternative approximate tests for a split-plot design. *Multivariate Behavioral Research*, *29*, 365–384.
- Anderson, T. W. (2003). *An introduction to multivariate analysis* (3rd ed.). New York: Wiley.
- Bathke, A. C., Harrar, S. W., & Madden, L. V. (2008). How to compare small multivariate samples using nonparametric tests. *Computational Statistics & Data Analysis*, *52*, 4951–4965.
- Blöte, A. W., Van der Burg, E., & Klein, A. S. (2001). Students' flexibility in solving two-digit addition and subtraction problems: Instruction effects. *Journal of Educational Psychology*, *93*, 627–638.
- Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effects of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, *25*, 290–403.
- Bradley, J. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152.
- Brunner, E., Dette, H., & Munk, A. (1997). Box-type approximations in factorial designs. *Journal of the American Statistical Association*, *92*, 1494–1503.
- Brunner, E., Munzel, U., & Puri, M. L. (2002). The multivariate nonparametric Behrens–Fisher problem. *Journal of Statistical Planning & Inference*, *108*, 37–53.
- Coombs, W. T., & Algina, J. (1996). New test statistics for MANOVA/descriptive discriminant analysis. *Educational and Psychological Measurement*, *56*, 382–402.
- Crowell, S. E., Beauchaine, T. P., McCauley, E., Smith, C. J., Vasilev, C. A., & Stevens, A. L. (2008). Parent–child interactions, peripheral serotonin, and self-inflicted injury in adolescents. *Journal of Consulting and Clinical Psychology*, *76*, 15–21.
- Farivar, S. S., Cunningham, W. E., & Hays, R. D. (2007). Correlated physical and mental health summary scores for the SF-36 and SF-12 Health Survey, V. I. *Health and Quality of Life Outcomes*, *5*, 54.
- Finch, W. H., & Davenport, T. (2009). Performance of Monte Carlo permutation and approximate tests for multivariate means comparisons with small sample sizes when parametric assumptions are violated. *Methodology*, *5*, 60–70.
- Gamage, J., Mathew, T., & Weerahandi, S. (2004). Generalized *p*-values and generalized confidence regions for the multivariate Behrens–Fisher problem and MANOVA. *Journal of Multivariate Analysis*, *88*, 177–189.
- Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, *40*, 177–190.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis: A global perspective* (7th ed.). Upper Saddle River: Pearson Prentice Hall.
- Hakstian, A. R., Roed, J. C., & Lind, J. C. (1979). Two-sample T^2 procedure and the assumption of homogeneous covariance matrices. *Psychological Bulletin*, *86*, 1255–1263.
- Holtzer, R., Friedman, R., Lipton, R. B., Katz, M., Xue, X., & Verghese, J. (2007). The relationship between specific cognitive functions and falls in aging. *Neuropsychology*, *21*, 540–548.
- Ito, P. K. (1980). Robustness of ANOVA and MANOVA test procedures. In P. R. Krishnaiah (Ed.), *Handbook of statistics* (Analysis of variance, Vol. 1, pp. 199–236). Amsterdam: North-Holland.
- James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*, *38*, 324–329.
- Jamshidian, M., & Schott, J. R. (2007). Testing equality of covariance matrices when data are incomplete. *Computational Statistics & Data Analysis*, *51*, 4227–4239.
- 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.
- Johnson, R. A., & Wichern, D. W. (2008). *Applied multivariate statistical analysis* (6th ed.). Upper Saddle River: Pearson Prentice Hall.
- Jonas, K., & Sassenberg, K. (2006). Knowing how to react: Automatic response priming from social categories. *Journal of Personality and Social Psychology*, *90*, 709–721.
- Kenward, M. G. (2001). Comments on “Parametric modelling of growth curve data: an overview”, by D. L. Zimmerman & V. Núñez–Antón. *TEST*, *10*, 1–73.
- Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983–997.
- Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, *53*, 2583–2595.
- Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods*, *13*, 110–129.
- Keselman, H. J., Carriere, K. C., & Lix, L. M. (1995). Robust and powerful nonorthogonal analyses. *Psychometrika*, *60*, 395–418.
- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*, 350–386.
- Krishnamoorthy, K., & Lu, F. (2010). A parametric bootstrap solution to the MANOVA under heteroscedasticity. *Journal of Statistical Computation and Simulation*, *80*, 873–887.
- Lindemann, O., & Bekkering, H. (2009). Object manipulation and motion perception: Evidence of an influence of action planning on visual processing. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1062–1071.
- Lix, L. M., Algina, J., & Keselman, H. J. (2003). Analysing multivariate repeated measures designs: A comparison of two approximate degrees of freedom procedures. *Multivariate Behavioral Research*, *38*, 403–431.
- Luh, W. M., & Guo, J. H. (2004). Improved robust test statistic based on trimmed means and Hall's transformation for two-way ANOVA models under non-normality. *Journal of Applied Statistics*, *31*, 623–643.
- Maas, C. J. M., & Snijders, T. B. (2003). The multilevel approach to repeated measures for complete and incomplete data. *Quality & Quantity*, *37*, 71–89.
- Mardia, K. V. (1971). The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika*, *58*, 105–121.
- Mathai, A. M., & Provost, S. B. (1992). *Quadratic forms in random variables: Theory and applications*. New York: Dekker.
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah: Erlbaum.
- Mehrotra, D. V. (1997). Improving the Brown–Forsythe solution to the generalized Behrens–Fisher problem. *Communications in Statistics–Simulation & Computation*, *26*, 1139–1145.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *92*, 778–785.
- Naik, D. N. (2003). Diagnostic methods for univariate and multivariate normal data. In R. Khattree & C. R. Rao (Eds.), *Handbook of statistics* (Statistics in industry, Vol. 22, pp. 957–993). New York: Elsevier Science.
- Nel, D. G. (1997). Tests for equality of parameter matrices in two multivariate linear models. *Journal of Multivariate Analysis*, *61*, 29–37.

- Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, *69*, 894–908.
- SAS Institute Inc. (2008). *SAS/STAT® 9.2 user's guide*. Cary: SAS Institute Inc.
- Schabenberger, O. (2007). Growing Up Fast: SAS® 9.2 Enhancements to the GLIMMIX Procedure. In *Proceedings of the 2007 SAS Global Forum*. Cary, NC: SAS Institute Inc.
- Sheehan-Holt, J. K. (1998). MANOVA simultaneous test procedures: The power and robustness of restricted multivariate contrasts. *Educational and Psychological Measurement*, *58*, 861–881.
- Vallejo, G., Arnau, J., & Ato, M. (2007). Comparative robustness of recent methods for the analysis of multivariate repeated measures designs. *Educational and Psychological Measurement*, *67*, 410–432.
- Vallejo, G., & Ato, M. (2006). Modified Brown–Forsythe procedure for testing interaction effects in split-plot designs. *Multivariate Behavior Research*, *41*, 549–578.
- Vallejo, G., Fernández, P., & Livacic-Rojas, P. (2010a). Analysis of unbalanced factorial designs with heteroscedastic data. *Journal of Statistical Computation and Simulation*, *80*, 75–88.
- Vallejo, G., Ato, M., & Fernández, P. (2010b). A robust approach for analyzing unbalanced factorial designs with fixed levels. *Behavior Research Methods*, *42*, 607–617.
- Vallejo, G., Fernández, M. P., Livacic-Rojas, P. E., & Tuero-Herrero, E. (2011). Selecting the best unbalanced repeated measures model. *Behavior Research Methods*, *43*, 18–36.
- Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330–336.
- Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). New York: Elsevier.