An Examination of the Robustness of the Empirical Bayes and Other Approaches for Testing Main and Interaction Effects in Repeated Measures Designs

by

H.J. Keselman, Rhonda K. Kowalchuk
University of Manitoba

and

Robert J. Boik
Montana State University

Abstract

Boik (1997) presented an empirical Bayes (EB) approach to the analysis of repeated measurements. The EB approach is a blend of the conventional univariate and multivariate approaches. Specifically, in the EB approach, the underlying covariance matrix is estimated by a weighted sum of the univariate and multivariate estimators. In addition to demonstrating that his approach controls test size and frequently is more powerful than either the ϵ-adjusted univariate or multivariate approaches, Boik showed how conventional multivariate software can be used to conduct EB analyses. Our investigation examined the Type I error properties of the EB approach when its derivational assumptions were not satisfied as well as when other factors known to affect the conventional tests of significance were varied. For comparative purposes we also investigated the procedures presented by Huynh (1978) and Keselman, Carriere and Lix (1993), procedures designed for non spherical data and covariance heterogeneity, as well as an adjusted univariate and multivariate test statistic. Our results indicate that when the response variable is normally distributed and group sizes are equal the EB approach was robust to violations of its derivational assumptions and therefore is recommended due to the power findings reported by Boik (1997). However, we also found that the EB approach, as well as the adjusted univariate and multivariate procedures, were prone to depressed or elevated rates of Type I error when data were nonnormally distributed and covariance matrices and group sizes were either positively or negatively paired with one another. On the other hand, the Huynh and Keselman et al. procedures were generally robust to these same pairings of covariance matrices and group sizes.

An Examination of the Robustness of the Empirical Bayes and Other Approaches for Testing Main and Interaction Effects in Repeated Measures Designs

The effects (e.g., main and interaction) that can be tested in repeated measures designs are typically based on the usual Gaussian linear model which can be written as

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

where $\mathbf{Y}$ is an $N \times t$ matrix of observations from N subjects, each observed on t occasions, $\mathbf{X}$ is an $N \times p$ matrix that codes for between-subjects effects where rank $(\mathbf{X}) = r \leq p$, $\mathbf{B}$ is a $p \times t$ matrix of unknown regression coefficients, and $\mathbf{R}$ is an $N \times t$ matrix of random errors. The rows of $\mathbf{R}$ are assumed to be *iid* as $\mathcal{N}_{\mathrm{t}}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a $t \times t$ positive definite covariance matrix.

Inferences about the linear functions of the regression coefficients are generally of interest. The linear functions of interest often can be represented as follows:

$$\mathbf{\Psi} = \mathbf{L}'\mathbf{BC}, \tag{2}$$

where $\mathbf{L}$ is a $p \times s$ contrast matrix with rank s for the between-subjects variable and $\mathbf{C}$ is a $t \times q$ orthonormalized contrast matrix for the repeated measures variable, where $q \leq t - 1$.

To test H_0: $\mathbf{\Psi} = \mathbf{\Psi}_0$ versus H_{a}: $\mathbf{\Psi} \neq \mathbf{\Psi}_0$, one first estimates $\mathbf{\Psi}$ via $\widehat{\mathbf{\Psi}} = \mathbf{L}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{YC}$, where under model (1) the distribution of $\widehat{\mathbf{\Psi}}$ is $\mathrm{vec}(\widehat{\mathbf{\Psi}}) \sim \mathcal{N}_{qs}[\mathrm{vec}(\mathbf{\Psi}), \mathbf{C}'\mathbf{\Sigma}\mathbf{C} \otimes \mathbf{L}'(\mathbf{X}'\mathbf{X})^{-}\ \mathbf{L}]$ [We use the notation $(\bullet)^{-}$ to represent any generalized inverse.]. Note that inferences about $\mathbf{\Psi}$ depend on $\mathbf{\Sigma}$ through

$$\mathbf{\Phi} = \mathbf{C}'\mathbf{\Sigma}\mathbf{C}. \tag{3}$$

The various approaches to the analysis of repeated measurements differ according to how they model $\mathbf{\Phi}$. The multivariate model places no constraints on $\mathbf{\Phi}$ other than that it must be positive definite. For this model the uniformly minimum variance unbiased estimator (UMVUE) of $\mathbf{\Phi}$ is

$$\widehat{\mathbf{\Phi}}_{\mathbf{MV}} = m^{-1}\mathbf{E}, \tag{4}$$

where $\mathbf{E} = \mathbf{C}'\mathbf{Y}'[\mathbf{I}_{\mathrm{N}} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}']\mathbf{YC}$ and $m \equiv N - r$.

In the univariate approach, on the other hand, $\mathbf{\Phi}$ is assumed to be spherical. That is,

$$\mathbf{\Phi} = \sigma^2 \mathbf{I}_q \tag{5}$$

and the UMVUE is

$$\widehat{\mathbf{\Phi}} = \widehat{\sigma}^2 \mathbf{I}_q, \tag{6}$$

where $\widehat{\sigma}^2 = \text{trace}(\mathbf{E})/(mq)$.

The adjusted df (ϵ-adjusted) univariate approach to the analysis of repeated measurements can be described as a hybrid between the conventional univariate and multivariate approaches. Specifically, the univariate approach requires that the underlying covariance matrix satisfy sphericity whereas the multivariate approach imposes no constraints on the covariance matrix. The ϵ-adjusted approach uses the usual *F* test with adjusted df. The approximation is based on Box's (1954) finding that the usual univariate *F* test statistic is approximately distributed as an $F(\epsilon qs, \epsilon mq)$ random variable, where

$$\epsilon = \frac{[\text{trace }(\mathbf{\Phi})]^2}{q \text{ trace }(\mathbf{\Phi}^2)}\ , \tag{7}$$

when sphericity is not satisfied. The ϵ-adjusted approach presumes that departures from sphericity are expected, but they are not expected to be extreme. The univariate estimator of the covariance matrix is retained and the sampling distribution of the statistic is adjusted for moderate departures from sphericity.

The Empirical Bayes Approach

An alternative univariate-multivariate hybrid, namely an empirical Bayes (EB) approach, was introduced by Boik (1997). The EB approach does not require sphericity for any specific covariance matrix. Rather, the EB approach requires that the average covariance matrix (averaged over all experiments) satisfies sphericity. This assumption is called second-stage sphericity. In the EB approach, the covariance matrix is estimated as a linear combination of the conventional univariate and multivariate estimators.

The EB approach to the analysis of repeated measurements requires that the data be sampled from a multivariate normal distribution and that the covariance matrix be sampled from a spherical inverted Wishart distribution. The approach uses a two-stage model. In the first stage a model similar to (1) is assumed except that only **YC** functions are modeled. Conditional on $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$, the first stage model is

$$\mathbf{YC} = \mathbf{X}\boldsymbol{\Theta} + \mathbf{U}, \tag{8}$$

where $\boldsymbol{\Theta} = \mathbf{BC}$ and $\mathbf{U} = \mathbf{RC}$. From Equation (1) it follows that the rows of $\mathbf{U}$ are *iid* $\mathcal{N}_q(\mathbf{0}, \boldsymbol{\Phi})$. In the second-stage, prior distributions on $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are assumed. Specifically, it is assumed that they are independently distributed, $\boldsymbol{\Theta}$ is uniformly distributed over a pq-dimensional space, and that $\boldsymbol{\Phi}$ follows a spherical inverted Wishart distribution. That is,

$$\boldsymbol{\Phi}^{-1} \sim W_q(f, \tau^{-1}\mathbf{I}), \tag{9}$$

It follows from equation (9) that

$$\mathrm{E}(\boldsymbol{\Phi}) = \boldsymbol{\sigma}^2\mathbf{I}_q, \text{ where } \sigma^2 = \frac{\tau}{f-q-1}. \tag{10}$$

That is, in the two-stage model, sphericity is satisfied on average though not on any particular experimental outcome. The hyperparameter f quantifies the prior belief about sphericity and it satisfies $q - 1 < f < \infty$. Small values of f reflect a belief that the departure from sphericity will be large while large values of f reflect a belief that departures from spherícty will be small.

Conventional multivariate software can be used to obtain EB analyses. Specifically, let $\mathbf{H}_b$ and $\mathbf{Q}_b$ represent the hypothesis and error matrices given by

$$\mathbf{H}_b = (\boldsymbol{\Psi}_\mathbf{0} - \widehat{\boldsymbol{\Psi}})'[\mathbf{L}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}]^{-1}(\boldsymbol{\Psi}_\mathbf{0} - \widehat{\boldsymbol{\Psi}}) \tag{11}$$

and

$$\mathbf{Q}_b = \tau\mathbf{I}_q + \mathbf{E}, \tag{12}$$

respectively. The eigenvalues of $\mathbf{Q}_b^{-1}\mathbf{H}_b$ have the same joint distribution as the eigenvalues of $\mathbf{E}_\mathrm{m}^{-1}\mathbf{H}_\mathrm{m}$ and $\mathbf{H}_\mathrm{m}$ and $\mathbf{E}_\mathrm{m}$ have independent Wishart distributions, namely,

$$\mathbf{H}_\mathrm{m} \sim W_q(s, \mathbf{I}_q) \qquad \mathbf{E}_\mathrm{m} \sim W_q(m + f, \mathbf{I}_q).$$

To obtain an empirical Bayes solution, one first estimates the hyperparameters f and τ from the observed data [see formulas (26), (29) and (31) in Boik, 1997]. Denote these estimators by $\widehat{f}$ and $\widehat{\tau}$ (our $\widehat{\tau}$ is $\widehat{\tau}$/c in Boik). Thus applied researchers can make inferences about $\boldsymbol{\Psi}$ by treating $\widehat{\mathbf{Q}}_b = \widehat{\tau}\,\mathbf{I}_q + \mathbf{E}$ as the error matrix with $m + \widehat{f}$ degrees of freedom and using $\mathbf{H}_b$ as the hypothesis matrix with s degrees of freedom with any of the conventional multivariate statistics.

Boik demonstrated, through Monte Carlo methods, that the EB approach adequately controls Type I error rate and that it is more powerful than both the ϵ-adjusted and multivariate procedures for many non-null mean configurations. Nonetheless, additional research is necessary to determine how robust the EB procedure is to violations of its derivational assumptions. That is, as indicated, the EB approach requires that the data be sampled from a multivariate normal distribution and that the covariance matrix be sampled from a spherical inverted Wishart distribution. Our investigation, therefore will examine the operating characteristics of the EB approach when the two assumptions are violated separately and jointly. Accordingly, this article examines the robustness of the EB approach.

Test Statistics

In addition to examining the EB approach to the analysis of repeated measurements we investigated, for comparative purposes, five other procedures; these included Huynh's (1978) Improved General Approximation (IGA) test, the adjusted df univariate test proposed by Quintana and Maxwell (1994), the nonpolar multivariate Welch-James (WJ) test (see Johansen, 1980 and Keselman, Carriere & Lix, 1993) and conventional multivariate tests. These procedures were selected for comparative purposes because they are either popular alternatives to the conventional univariate test (multivariate test, ϵ-adjusted test) or, based on prior literature, likely to be robust in cases were the EB approach may not (IGA, WJ).

The IGA test is a univariate test that adjusts the df of the usual F test to account for violations of multisample sphericity (see Algina, 1997, Algina & Oshima, 1994, 1995; Keselman & Algina, 1996; Huynh, 1978). WJ, on the other hand, is a multivariate statistic that does not require sphericity and allows for heterogeneity of the between-subjects covariance matrices by using a non pooled estimate of error and a sample estimate of df (see Keselman & Algina, 1996; Keselman, et al., 1993). The IGA and WJ tests have been shown to be relatively insensitive to violations of mutisample sphericity and nonnormality even in unbalanced designs (see Algina & Keselman, 1997; Keselman, Algina, Kowalchuk & Wolfinger, 1997, 1999). The ϵ-adjusted df univariate test that we examined was proposed by Quintana and Maxwell (1994). With this test the adjustment is based on the adjustments due to Greenhouse and Geisser (1959) and Huynh and Feldt (1976), $\widehat{\epsilon}$ and $\widetilde{\epsilon}$, respectively. Specifically, their sample estimate of the unknown sphericity parameter is $\overline{\epsilon} = \frac{1}{2}(\widehat{\epsilon} + \widetilde{\epsilon})$, where

$$\widehat{\epsilon} = \frac{[\text{trace } (\widehat{\boldsymbol{\Phi}}_{\text{MV}})]^2}{q \text{ trace } (\widehat{\boldsymbol{\Phi}}^{\mathbf{2}}_{\text{MV}})}\ , \tag{13}$$

and $(\widehat{\boldsymbol{\Phi}}_{\text{MV}})$ is given in (4) and

$$\widetilde{\epsilon} = \min\left[\frac{(m+1)q\,\widehat{\epsilon}-2}{q(m-q\,\widehat{\epsilon})}, 1\right]. \tag{14}$$

Finally, we computed Hotelling's (1931) T^2 when examining the repeated measures main effect and the (a) the Hotelling (1951)-Lawley (1938) (HL) trace criterion, (b) the Pillai (1955)-Bartlett (1939) (P) trace statistic, and (c) Wilk's (1932) (W) likelihood ratio, when examining the interaction effect. These multivariate tests were computed in two ways, that is, they were based on either their conventional formulations or on the EB estimate of the covariance matrix.

Methods of the Simulation

The various approaches to the analysis of repeated measurements were examined in a between-subjects by within-subjects repeated measures design. There were three levels of the grouping variable and four levels of the within-subjects variable. Seven variables were examined in our simulation study.

The first two variables examined relate to one of the assumptions required for the EB approach. That is, as indicated, the EB approach requires that the covariance matrix be sampled from a spherical inverted Wishart distribution. The following sampling schema was used in order to simulate second-stage sphericity. Random positive definite covariance matrices were generated in the following manner. Let $\mathbf{Z}$ be an $f \times t$ random matrix in which the entries are $\mathcal{N}(0, 1)$ variables and let $\mathbf{V}$ be the matrix square root of $\mathbf{C'\Sigma C}$, that is, a fixed nonsingular matrix of size $q \times q$, where $\mathbf{\Sigma}$ is a covariance matrix with known epsilon and $\mathbf{C}$ is as previously defined. Then $\mathbf{\Phi} = \mathbf{(V'Z'ZV)}^{-1}$ has an inverted Wishart distribution. If $\mathbf{C'\Sigma C} = \sigma^2\mathbf{I},$ and $\mathbf{Z}$ is Gaussian, then $\mathbf{\Phi} = \mathbf{(V'Z'ZV)}^{-1}$ follows a spherical inverted Wishart distribution (See Boik, 1997). The inverted Wishart assumption can be violated in two ways. If $\mathbf{C'\Sigma C} \neq \sigma^2\mathbf{I},$ and $\mathbf{Z}$ is Gaussian, then $\mathbf{\Phi} = \mathbf{(V'Z'ZV)}^{-1}$follows a non-spherical inverted Wishart distribution. If $\mathbf{Z}$ is not Gaussian, then $\mathbf{\Phi} = \mathbf{(V'Z'ZV)}^{-1}$does not follow an inverted Wishart distribution.

To examine $\mathbf{C'\Sigma C} \neq \sigma^2\mathbf{I},$ we varied the value of sphericity (ϵ) of the population covariance matrix ($\mathbf{\Sigma}$). Specifically, we estimated Type I error rates for the procedures when $\epsilon = 1.00$, .75 and .40. Thus, when $\epsilon = 1.00$ our sampling schema conforms to the requirements of second-stage sphercity. Having $\epsilon = .75$ and .40 will enable us to examine the operating characteristics of the EB approach when, on average, sphericity is not satisfied. The element values of the .75 and .40 covariance matrices (i.e., $\mathbf{\Sigma}$) can be found in Keselman and Keselman (1990).

To investigate the second manner in which second-stage sphericity can be violated, we generated positive definite covariance matrices that do not follow an inverted Wishart distribution by obtaining **Z** from nonnormal continuous distributions. In particular, we obtained **Z** from either a *t* distribution (df = 3) or from a lognormal distribution [$\mathbf{Z} = \exp(\mathbf{R}) - \exp(1/2)$, where $\mathbf{R} \sim N(0,1)$]. Sample data were then generated, for each simulation, from multivariate distributions having covariance matrix $\mathbf{\Phi}$. Pseudorandom observation vectors $\mathbf{Y}'_{\mathrm{ij}} = [\mathbf{Y}_{\mathrm{ij1}} \ldots \mathbf{Y}_{\mathrm{ijq}}]$ with mean vector $\boldsymbol{\mu}'_{\mathrm{j}} = [\boldsymbol{\mu}_{\mathrm{j1}} \ldots \boldsymbol{\mu}_{\mathrm{jq}}]$ and covariance matrix $\mathbf{\Phi}_{\mathrm{j}}$ were obtained from a *q*-variate normal distribution. The observation vectors were obtained by a triangular decomposition of $\mathbf{\Phi}_{\mathrm{j}}$; that is, $\mathbf{Y}_{\mathrm{ij}} = \boldsymbol{\mu}_{\mathrm{j}} + \mathbf{L} * \mathbf{ZN}_{\mathrm{ij}}$, where **L** is a lower triangular matrix satisfying the equality $\mathbf{\Phi}_{\mathrm{j}} = \mathbf{LL}'$ and $\mathbf{ZN}_{\mathrm{ij}}$ is an independent normally distributed unit vector obtained by the RANNOR function (SAS, 1989). The nonnormal χ^2_3 data were created by summing the squared values of three N(0,1) variates and standardizing the resulting sum.

The remaining six factors examined in our study were: (a) the value of *f*, (b) total sample size, (c) equality/inequality of the between-subjects group sizes, (d) equality/inequality of the group covariance matrices, (e) pairing of the covariance matrices and group sizes, and (f) distributional form of the response variable. It is important to note that when covariance matrices were equal across groups our results will be relevant to what can be expected for the test of a repeated measures variable in a simple repeated measures design, that is a design containing no between-subjects grouping variables (Kirk, 1995; Maxwell & Delaney, 1990; Rogan et al., 1979).

The size of f could affect the performance of the EB procedure and accordingly was varied. In particular, we set $f = 1, 2, 3, 5(q)$.

Sample size affects the relative power of the ϵ-adjusted univariate and multivariate tests. That is, when sample size is small, the power of the multivariate test will be low and likely less than the adjusted univariate test because of the imprecision in estimating all of the variances and covariances (Boik, 1997). As sample size increases, however, the power of the multivariate test improves and can be greater than the power of the ϵ-adjusted test. Clearly sample size affects the EB approach as well. Consequently, in our investigation we set sample size at three values: $N = 18, 30, 45$. These sample sizes were chosen, in part, because according to the survey of the educational and psychological literature reported by Kowalchuk, Lix and Keselman (1996), over 30% (50%) of the repeated measures articles they reviewed had fewer that 30 (60) observations in simple (mixed) designs. Additionally, we investigated two small sample size conditions, namely $N = 6$ and $N = 9$. We investigated these small sample size cases because we felt the EB approach might prove superior in these cases compared to the ϵ-adjusted and multivariate procedures.

In addition, to varying the total sample size, we also investigated the effect of group size balance/imbalance. Balance/imbalance was varied because the effect of other conditions (e.g., covariance heterogeneity) are known to be exacerbated by group imbalance. Furthermore, a recent survey indicates that imbalance is the norm and not the exception in behavioral science research (see Keselman et al., 1998). For each value of N two cases of imbalance were investigated, where the second case for each N the sizes were more disparate: (a) 4, 6, 8 and 3, 6, 9 for $N = 18$, (b) 8, 10, 12 and 7, 10, 13 for $N = 30$, and (c) 13, 15, 17 and 12, 15, 18 for $N = 45$.

Previous research pertaining to the ϵ-adjusted univariate and multivariate tests indicated that between-group covariance heterogeneity affects rates of Type I error, particularly when group sizes are unequal. Accordingly, we varied the equality/inequality of the group covariance matrices. When unequal, the matrices were multiples of one another, namely $\mathbf{\Phi}_1 = \frac{1}{3}\mathbf{\Phi}_2$, and $\mathbf{\Phi}_3 = \frac{5}{3}\mathbf{\Phi}_2$. This degree and type of covariance heterogeneity was selected because it has been used by others in research involving the analysis of repeated measurements (see e.g., Algina & Keselman, 1997; Keselman, et al., 1997, 1999; Keselman et al., 1993).

Six pairings of covariance matrices and group sizes ($\mathbf{\Phi}_j$& n_j) were investigated: (a) equal n_j; equal $\mathbf{\Phi}_j$, (b) equal n_j ; unequal $\mathbf{\Phi}_j$, (c/c′) unequal n_j ; unequal $\mathbf{\Phi}_j$ (positively paired), and (d/d′) unequal n_j ; unequal $\mathbf{\Phi}_j$ (negatively paired). The c′/d′ condition refers to the more disparate unequal group sizes case while the c/d condition designates the less disparate unequal group sizes case. A positive pairing results when the largest group size is associated with the covariance matrix containing the largest element values whereas a negative pairing results when the largest group size is associated with the covariance matrix with the smallest element values.

The last variable investigated was the distributional shape of the response variable. In particular, the form of the variable was either multivariate normal or $\chi^2_{(3)}$ distributed (see Keselman et al., 1993). The effect of nonnormality on the EB approach is predictable since the EB procedure operates like the conventional multivariate procedures when sample size is large and more like the conventional univariate procedure as sample size decreases and/or departure from sphericity decreases. Nonetheless, for completeness, we will table a few exemplars to verify this observation.

Type I error rates were collected over 5,000 replications per investigated condition. We believe that this number of replications results in stable estimates of the Type I error rates.

Results

To evaluate the particular conditions under which a test was insensitive to assumption violations, Bradley's (1978) liberal criterion of robustness was employed.

According to this criterion, in order for a test to be considered robust, its empirical rate of Type I error ($\widehat{\alpha}$) must be contained in the interval $0.5\alpha \leq \widehat{\alpha} \leq 1.5\alpha$. Therefore, for the five percent level of significance used in this study, a test was considered robust in a particular condition if its empirical rate of Type I error fell within the interval $.025 \leq \widehat{\alpha} \leq .075$. Correspondingly, a test was considered to be non robust if, for a particular condition, its Type I error rate was not contained in this interval. In the tables, bolded entries are used to denote these latter values. We chose this criterion since we feel that it provides a reasonable standard by which to judge robustness. That is, in our opinion, applied researchers should be comfortable working with a procedure that controls the rate of Type I error within these bounds, if the procedure limits the rate across a wide range of assumption violation conditions. Nonetheless, there is no one universal standard by which tests are judged to be robust or not and thus with other standards different interpretations of the results are possible.

For space considerations, we do not present results for the usual multivariate tests of the repeated measures interaction effect; these results are predictable from results that are tabled. Instead, we present all three EB interaction test results because this represents new information. Furthermore, from a preliminary analysis of the data we found that the rates of error were very similar when $\epsilon = 1.00$ and 0.75; accordingly, we do not always table rates of Type I error when $\epsilon = .75$.

Z Normally Distributed

Table 1 contains empirical rates of Type I error (%) for the six combinations (a-d') of $\mathbf{\Phi}_j$& n_j investigated when $N = 18$ for $f = 3$ and 15 when **Z** was obtained from a normal distribution (WJ rates are not tabled because sample sizes were smaller than those recommended by Keselman et al., 1993).[1] Apparent from this table is that the empirical values generally did not differ substantially across these two values of *f* investigated; furthermore, rates of Type I error were not consistently larger for either value of *f*. The remaining values of f that were investigated ($f = 6$ and 9) resulted in rates that were similar to those reported in Table 1.

The empirical values tabled for condition (a) (homogeneity of between-subjects covariance matrices) when $\epsilon = 1.0$ indicate how the EB approach to the analysis of repeated measurements (in either a simple or mixed design) performs under second-stage sphericity. Tabled values were all within the Bradley (1978) interval. Also evident from Table 1 is that the EB tests' Type I error rates were, with the exception of the HL test of the interaction effect, also within Bradley's interval for conditions (a) and (b) (equal group sizes) when $\epsilon = .40$ (and for the non tabled $\epsilon = .75$ values).

Further examination of Table 1 indicates that all procedures were affected by violations of multisample sphericity in at least one of the conditions investigated. The IGA test resulted in liberal values under condition d (7.86% for the test of the interaction effect when $f = 3$) and d' (e.g., 9.02% for the test of the main effect when $f = 3$) (negative pairing of covariance matrices and group sizes). As expected, the adjusted df univariate and multivariate tests were generally not robust to combinations of unequal covariance matrices and unequal group sizes (conditions c/c', d/d'). In particular, the tests were conservative for positive pairings of covariance matrices and group sizes and liberal for negative pairings of covariance matrices and group sizes; conservative values were as low as .74% (for the $\overline{\epsilon}$ test of the main effect when $f = 15$) while liberal values were as large as 19.78% (for the T^2 test of the main effect when $f = 15$). Not surprisingly, the EB tests were likewise affected by covariance heterogeneity when the design was unbalanced. That is, rates were conservative (e.g., .82% for the test of the main effect when $f = 3$) for positive pairings of covariance matrices and group sizes and liberal (e.g., 22.42% for the test of the interaction effect when $f = 15$) for negative pairings of covariance matrices and group sizes. The HL version of EB was also frequently liberal when covariance matrices were unequal but group sizes were equal [condition (b)]. Of the three EB tests of the interaction effect, the HL values were most liberal, followed by

W and then P. The preceding findings were very similar across the three values of ϵ investigated.

Since rates of Type I error do not appear to be affected by f, the remaining tabled values were based on one of our intermediate investigated values, i.e., $f = 6$. The rates in Table 2 are for the sample size cases of $N = 30$ and $N = 45$. Because the $N = 30$ unequal sample size conditions (8, 10, 12 and 7, 10, 13) are close, though still smaller, than the recommended values prescribed by Keselman et al. (1993), WJ values are tabled along with the other investigated tests.

Once again when sphercity is satisfied and covariance matrices are equal across groups, the EB procedure maintains effective Type I error control. Furthermore, control is maintained, when group sizes were equal (conditions (a) and (b)), even when $\epsilon = .40$ (and for the non tabled $\epsilon = .75$ values). As we found when $N = 18$, the adjusted df univariate, multivariate and EB tests were prone to depressed and inflated rates of error when covariance matrices and group sizes were unequal. The rates however, were not quite as distorted for these larger sample size cases. In particular, the rates when $N = 30$ were elevated, ranging in value around 10%-12%. On the other hand, the IGA test was always robust to violations of multisample sphericity, while WJ was robust except for the test of the interaction effect under condition d and d' (Remember the $N = 30$ sample sizes were smaller than those recommended by Keselman et al.). When $N = 45$, the adjusted df, multivariate and EB tests were once again prone to elevated rates of Type I error for negative pairings, though rates were rarely conservative for positive pairings of covariance matrices and group sizes. The EB rates were in the 8% to 11% range. On the other hand, the rates for the IGA and WJ tests were always well controlled.

Z Nonnormally Distributed ($t = 3$)

As previously indicated, if **Z** is not Gaussian, then $\mathbf{\Phi} = (\mathbf{V'Z'ZV})^{-1}$does not follow an inverted Wishart distribution. The results in Table 3 for $f = 3$, $N = 18$ and $f = 6$, $N = 45$ indicate that when the rows of **Z** were distributed as t variates (with three df) the effect on Type I error rates was minimal. For the $f = 3$, $N = 18$ data, when **Z** was obtained from a normal distribution, 54 of the tabled empirical values fell outside of Bradley's (1978) interval while the number of estimates that were outside the interval was 55 when **Z** was obtained from a *t* distribution. For $f = 6$, $N = 45$ data, the corresponding number of non robust empirical values was values was 22 and 27, respectively.

Z Nonnormally Distributed (Lognormal)

Table 4 contains empirical rates of Type I error when **Z** was lognormally distributed for various combinations of f, N and ϵ. The effect of obtaining **Z** from a non-symmetric distribution was similar to the effect of obtaining **Z** from the symmetric *t* and normal distributions. That is, the EB approach was robust under conditions (a) and (b), (group sizes are equal) and generally non robust under conditions (c), (d), (c'), and (d') (group sizes are unequal). Specifically, for the test of the repeated measures main effect, the non robust rates of error ranged from 1.08% to 19.06%, whereas for the interaction effect test, the rates ranged from 1.08% to 21.68%. Also, as was found with the normal and *t* data, the IGA test was robust except in condition (d) (7.64% for the interaction test) and (d') (8.96% for the main effect test and 9.68% for the interaction test) when $N = 18$. As well, the WJ test was robust except in condition (d') (8.30% for interaction test) for $N = 30$. The results for the $\overline{\epsilon}$-adjusted test were also similar to prior results.

Y Nonnormally Distributed/**Z** Nonnormally Distributed (Lognormal)

Table 4 also contains empirical rates of Type I error when the response variable was χ^2_3 distributed, rather than normally distributed. Specifically, as a point of comparison, we have tabled empirical values for the same combinations of f, N and ϵ investigated when the response variable was normally distributed, namely when $f = 3$, $N = 18$, $\epsilon = 1.00$ and $f = 6$, $N = 45$, $\epsilon = 0.40$. The effect of the response variable being nonnormally distributed resulted, on average, in the liberal values being slightly elevated compared to their normal distribution counterparts. That is, when $f = 3$, $N = 18$, $\epsilon = 1.00$, the average liberal Type I error rate was 14.2% for nonnormal data compared to 14% for normally distributed data. The corresponding values for $f = 6$, $N = 45$, $\epsilon = 0.40$ were 9.7% and 9.4%, respectively.

Small Sample Behavior

We also investigated the Type I error behaviour of the EB, $\overline{\epsilon}$-adjusted and multivariate procedures when sample sizes were small, i.e., $N = 6$ and $N = 9$, because the EB approach may perform better than its competitors in such cases. We only table conditions (a) and (b) because the results for conditions involving unequal group sizes are predictable from the results previously presented. Table 5 contains rates for normally distributed data while Table 6 presents rates for χ_3^2 distributed data. Because the IGA and WJ procedures require larger sample sizes they were not investigated for these sample size cases.

Normally Distributed Data. For the main effect test, the $\overline{\epsilon}$-adjusted and multivariate procedures always had rates of error within Bradley's (1978) liberal interval. The empirical EB values, with one exception (7.66%) were also within Bradley's interval. For the interaction effect, only the $\overline{\epsilon}$-adjusted and EB Pillai tests resulted in robust values [condition (b) has unequal covariance matrices]. The Hotelling-Lawley and Wilks EB tests frequently had liberal rates of Type I error (e.g., in excess of 8%).

χ_3^2 Distributed Data. When the response variable was not normally distributed main effect rates of Type I error were much larger than the values enumerated in Table 5. Thus, most of the main effect values fell outside of Bradley's (1978) liberal interval, attaining values in the 8%-12% range. On the other hand, the rates for the test of the interaction effect were very similar to their Table 5 counterparts, meaning that by in large, the procedures, excluding HL, were robust to nonnormal.

Discussion

In our paper we compared one of the newest approaches to the analysis of repeated measurements, the empirical Bayes approach presented by Boik (1997), to procedures that are frequently used by researchers to analyze repeated measures hypotheses (an adjusted df univariate approach and multivariate statistics) and to procedures that have been reported to be generally robust to non sphericity, covariance heterogeneity, and nonnormality [Huynh's (1978) IGA test and Johansen's (1980) WJ test]. The approach involves a blending of the univariate and multivariate approaches to the analysis of repeated measurements. The EB approach requires that the data be sampled from a multivariate normal distribution and that the covariance matrix be sampled from a spherical inverted Wishart distribution. Our investigation, examined the Type I error operating characteristics of the EB approach when the two assumptions were violated separately and jointly.

Our findings with respect to the EB approach varied with whether the response variable was normally distributed, whether group sizes were equal, and whether the approach was being used to analyze main or interaction effects. When the response variable was normally distributed, the EB approach for testing main and interaction effects was robust even though Φ was not distributed as a spherical inverted Wishart variable when group sizes were equal. This finding held even when covariance matrices were unequal. However, when the response variable was not normally distributed, the EB approach tended to be non robust In particular, main effect Type I error rates were liberal when sample size was small ($N = 18$) even when the data were spherical. That is, equal group sizes did not guarantee robustness. For larger sample sizes ($N = 45$), the EB procedure was not robust when covariance matrices were unequal even though group sizes were equal. On the other hand, the EB approach generally resulted in a robust test when applied to the interaction effect, particularly when the Pillai-Bartlett and Wilks criteria were adopted. Thus, we could only recommend the EB approach when data are known to be normally distributed and covariance matrices are homogeneous (or group sizes are equal).

In our investigation we also compared the empirical Bayes approach with an adjusted df univariate test (due to Quintana & Maxwell, 1994), a multivariate test, the

IGA procedure due to Huynh (1978) and the WJ procedure presented by Keselman et al. (1993). Furthermore, the Monte Carlo investigation also varied equality/inequality of the between-subjects group sizes, equality/inequality of the group covariance matrices, and pairings of the covariance matrices and group sizes.

We examined combinations of covariance homogeneity/heterogeneity and group size homogeneity/heterogeneity because we believed the EB approach might be also adversely affected by the same factors which affect the validity of the approaches which comprise the EB approach, namely covariance heterogeneity when occurring in unbalanced designs. Since unbalanced designs are the norm and not the exception in behavioural science research according to a recent survey by Keselman et al. (1998), the effects that covariance heterogeneity might have on this newest of approaches to the analysis of repeated measurements should be of interest to behavioural science researchers. In particular, since the EB approach was not designed for heterogeneity it was reasonable to assume that heterogeneity when present would distort rates of Type I error.

We found, as expected, that the empirical Type I error rates of the adjusted df univariate and multivariate tests were adversely affected by heterogeneity of covariance matrices when group sizes were unequal. In particular, the rates were either depressed or elevated depending on whether covariance matrices and group sizes were positively or negatively paired. Not surprisingly therefore, the EB tests were similarly affected. On the other hand, the IGA and WJ tests were generally robust to these pairings of unequal covariances and unequal group sizes except when N was small (i.e., $N = 18$) and thus the smallest of the unequal group sizes was particularly small (e.g., 3). Though we did not table WJ values when sample sizes were small, results reported by Keselman et al., (1993) and Keselman, Keselman and Lix (1995) indicate that the procedure is robust to nonnormality and covariance heterogeneity as long as sample sizes conform to the prescriptions given by Keselman et al. (1993).

Therefore, when group sizes are unequal and covariance matrices may be heterogeneous, we recommend either the IGA or WJ approach. Results presented here as well as elsewhere indicate that these procedures are generally robust to the effects of nonsphericity and covariance heterogeneity (see Algina & Keselman, 1997; Keselman et al., 1997, 1999; Keselman et al., 1993). Based on the power results presented by Algina and Keselman (1998) one can expect a more powerful test of a repeated measures effect with the WJ test. However, to obtain a robust WJ test, sample sizes need to be larger than is the case with the IGA test. Thus, when researchers do not have the requisite sample sizes prescribed by Keselman et al. (1993) and Algina and Keselman (1997), the IGA test should be used.

As a postscript we point out that in making the preceding recommendations we have taken into account the literature regarding mixed-model analyses for repeated measurements. That is, the mixed-model approach to the analysis of repeated measurements is advocated by its proponents because it allows users to model the correct covariance structure of their data, thereby, presumably, as previously indicated, resulting in more powerful tests of repeated measures effects. However, results reported by Keselman et al. (1997, 1999) indicate that the default and Satterthwaite *F* tests that are computed with the SAS (Littell, Milliken, Stroup & Wolfinger, 1996) mixed-model program (PROC MIXED) are also prone to depressed or inflated rates of Type I error when covariance matrices and group sizes are unequal and positively or negatively paired, respectively. Thus, these findings also played a role in our recommendations.

Footnotes

[1] According to Keselman et al. (1993), when data are normally distributed, to obtain a robust test of the repeated measures main effect hypothesis, the number of observations in the smallest of groups must be 2 to 3 times the number of repeated measurements minus one (i.e., $K - 1$); to obtain a robust test of the interaction, this number must be 3 or 4 times ($K - 1$). The corresponding values for nonnormally distributed data are 3 or 4 times one and 5 or 6 times one, respectively. Algina and Keselman (1997) determined that the sample size requirements enumerated by Keselman et al. (1993) generalize to larger repeated measures designs (i.e., 6 × 4 and 6 × 8 as opposed to the 3 × 4 and 3 × 8 designs investigated by Keselman et al., 1993) for the test of the main effect but that sample size requirements had to be larger in order to obtain a robust interaction test.

References

Algina, J. (1997). Generalization of improved general approximation tests to split-plot designs with multiple between-subjects factors and/or multiple within-subjects factors. British Journal of Mathematical and Statistical Psychology, 50, 243-252.

Algina, J. (1994). Some alternative approximate tests for a split plot design. Multivariate Behavioral Research, 29, 365-384.

Algina, J., & Keselman, H. J. (1998). A power comparison of the Welch-James and Improved General Approximation tests in the split-plot design. Journal of Educational and Behavioral Statistics, 23, 152-169.

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.

Algina, J., & Oshima, T. C. (1994). Type I error rates for Huynh's general approximation and improved general approximation tests. British Journal of Mathematical and Statistical Psychology, 47, 151-165.

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

Boik, R. J. (1997). Analysis of repeated measures under second-stage sphericity: An empirical Bayes approach. Journal of Educational and Behavioral Statistics, 22, 155-192.

Greenhouse, S. W. & Geisser, S. (1959). On methods in the analysis of profile data. Psychometrika, 24, 95-112.

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

Hotelling, H. (1951). “A Generalized t Test and Measure of Multivariate Dispersion,” In J. Neyman (Ed.). Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 2, 23-41.

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

Huynh, H. & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F distributions. Journal of the American Statistical Association, 65, 1582-1589.

Huynh, H. & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. Journal of Educational Statistics, 1, 69-82.

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. Biometrika, 38, 324-329.

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

Keselman, H. J. & Algina, J. (1996). In B Thompson's (Ed) The analysis of higher-order repeated measures designs. In Advances in social science methodology (Vol 4), Greenwich, Conn: JAI Press.

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., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1997). The analysis of repeated measurements with mixed-model Satterthwaite F tests. Unpublished manuscript.

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

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of Educational Researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. Review of Educational Research, 68(3), 350-386.

Keselman, H. J., Keselman, J. C., & Lix, L. M. (1995). The analysis of repeated measurements: Univariate tests, multivariate tests, or both? British Journal of Mathematical and Statistical Psychology, 48, 319-338.

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

Kirk, R. E. (1995). Experimental design: Procedures for the behavioral sciences (3rd ed). Belmont, CA: Brooks/Cole.

Lawley, D. N. (1938). A generalization of Fisher's z test. Biometrika, 30, 180-187, 467-469.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). SAS system for mixed models, Cary, NC: SAS Institute.

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

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

Quintana, S. M., & Maxwell, S. E. (1994). A Monte Carlo comparison of seven ϵ-adjustment procedures in repeated measures designs with small sample sizes. Journal of Educational Statistics, 19, 57-71.

Rogan, J. C., Keselman, H. J., & Mendoza, J. L. (1979). Analysis of repeated measurements. British Journal of Mathematical and Statistical Psychology, 32, 269-286.

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

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

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

Table 1. Empirical Type I Error Rates (%) (**Z** Normally Distributed)

є/C	Main Effect Test				Interaction Effect Test				
	IGA	$\bar{\epsilon}$	T^2	EB	IGA	$\bar{\epsilon}$	HL(EB)	P(EB)	W(EB)
f=3, *N*=18									
1.0/a	5.36	5.06	4.90	4.76	6.46	5.30	5.46	3.82	5.00
b	5.34	5.36	6.06	5.76	6.68	6.26	**7.92**	5.40	6.98
c	5.12	2.60	**2.42**	**2.14**	5.52	3.12	3.52	**1.88**	2.98
d	6.72	**10.48**	**13.72**	**12.90**	**7.86**	**11.16**	**15.10**	**11.22**	**13.66**
c’	4.84	**1.28**	**1.06**	**0.82**	5.40	**1.96**	**2.24**	**0.96**	**1.58**
d’	**8.46**	**14.10**	**19.48**	**19.02**	**9.24**	**14.72**	**21.64**	**17.76**	**20.38**
.40/a	5.02	5.04	4.98	5.04	6.28	5.08	5.58	3.48	5.02
b	5.18	5.48	6.78	6.30	5.70	5.66	7.34	4.64	6.40
c	4.72	**2.42**	**2.10**	**2.00**	5.88	3.66	3.40	**1.62**	2.70
d	7.40	**10.48**	**14.62**	**13.68**	7.40	**10.14**	**15.28**	**10.76**	**13.64**
c’	4.98	**1.68**	**1.32**	**1.34**	5.68	2.60	**2.26**	**1.06**	**1.74**
d’	**9.02**	**12.92**	**19.28**	**18.72**	**9.36**	**13.96**	**20.64**	**15.82**	**18.72**
f=15, *N*=18									
1.0/a	5.46	5.12	5.48	6.08	5.50	4.82	6.30	5.38	5.94
b	4.96	4.72	6.24	5.98	5.44	5.60	**7.64**	6.20	7.06
c	5.62	**2.06**	**2.28**	**2.16**	5.40	**2.28**	2.98	**1.98**	**2.42**
d	5.74	**10.72**	**14.24**	**13.76**	6.06	**12.30**	**16.34**	**13.72**	**15.20**
c’	4.82	**0.74**	**1.14**	**0.94**	6.00	**1.78**	**2.48**	**1.62**	**1.98**
d’	6.80	**16.38**	**18.94**	**20.08**	7.10	**17.10**	**22.42**	**19.64**	**21.50**
.40/a	5.14	5.16	4.96	4.76	6.12	5.16	5.14	3.64	4.78
b	5.02	5.44	6.44	5.94	6.86	6.72	**7.70**	4.62	6.64
c	5.00	2.64	2.78	**2.16**	5.94	3.76	2.98	**1.82**	2.50
d	6.66	**9.80**	**13.80**	**13.16**	**7.64**	**10.38**	**15.64**	**11.94**	**14.24**
c’	5.10	**1.94**	**1.16**	**0.96**	5.62	**2.20**	**2.20**	**1.10**	**1.72**
d’	**8.54**	**12.76**	**19.78**	**18.56**	**9.04**	**12.92**	**20.56**	**15.74**	**18.86**

Note: є/C-Value of the sphericity parameter/Pairing of covariance marices and group sizes condition; IGA-Huynh’s (1978) ImprovedGeneral Approximation Test,$\bar{\epsilon}$-EBAR (Quintana & Maxwell, 1994),T^2-Hotelling’s test,HL-Hotelling-Lawley Trace, P-Pillai test, W-Wilks test, EB-Empirical Bayes.

Table 2. Empirical Type I Error Rates (%) (**Z** Normally Distributed)

ϵ/C	Main Effect Test					Interaction Effect Test					
	IGA	WJ	$\bar{\epsilon}$	T^2	EB	IGA	WJ	$\bar{\epsilon}$	HL(EB)	P(EB)	W(EB)
f=6, *N*=30											
1.0/a	5.40	5.14	5.22	5.30	5.22	6.34	5.02	4.94	4.94	4.02	4.46
b	5.34	4.80	5.30	5.72	5.46	6.76	6.72	6.40	7.44	5.72	6.74
c	4.78	4.86	2.88	3.14	2.92	6.06	5.94	3.90	4.10	3.18	3.76
d	5.24	5.08	**7.88**	**10.10**	**9.74**	6.86	**8.18**	**9.66**	**11.64**	**9.92**	**10.78**
c’	4.78	4.06	**2.30**	**2.00**	**1.78**	6.06	5.70	2.90	3.20	**2.48**	2.88
d’	6.14	5.68	**10.52**	**12.68**	**12.50**	6.76	**8.46**	**11.04**	**13.80**	**11.54**	**12.86**
.40/a	5.10	4.90	5.12	5.26	4.90	6.08	5.26	4.76	4.64	3.82	4.22
b	4.84	5.10	4.98	5.98	5.78	6.54	6.40	6.32	6.98	5.40	6.28
c	4.64	4.60	3.20	3.28	2.84	6.86	6.00	4.70	4.40	3.18	3.76
d	5.06	4.82	7.14	**9.72**	**8.94**	6.56	7.58	**8.50**	**11.06**	**8.64**	**10.04**
c’	4.68	4.84	**2.36**	**2.12**	**1.90**	6.40	5.78	3.60	3.24	**2.18**	2.68
d’	6.00	5.70	**9.18**	**12.64**	**12.36**	6.76	**8.46**	**9.72**	**14.20**	**11.90**	**13.18**
f=6, *N*=45											
1.0/a	5.02	4.44	4.90	4.62	4.52	6.84	4.98	5.00	5.14	4.66	4.90
b	5.08	5.12	5.10	5.78	5.74	6.30	5.80	5.86	6.72	5.92	6.26
c	4.80	4.62	3.54	3.54	3.30	6.24	4.86	4.30	4.52	3.78	4.20
d	4.70	4.58	6.58	**7.60**	7.26	5.82	5.36	7.30	**8.84**	**7.94**	**8.38**
c’	4.68	4.28	2.64	**2.48**	**2.46**	6.50	5.36	3.92	3.88	3.26	3.54
d’	5.38	4.94	**8.28**	**9.74**	**9.60**	7.10	5.86	**9.62**	**10.92**	**9.86**	**10.54**
.40/a	5.18	4.66	5.12	4.78	4.78	6.82	4.80	4.96	4.82	4.12	4.62
b	5.60	5.40	5.70	6.12	5.74	6.24	5.50	5.74	6.84	5.62	6.40
c	4.78	4.88	3.70	3.62	3.40	6.54	5.58	4.96	4.50	3.72	4.10
d	4.62	4.40	6.24	7.40	6.98	5.80	5.44	6.56	**8.24**	7.28	**7.78**
c’	4.62	4.92	3.40	3.20	2.86	5.50	4.88	3.88	3.32	2.68	3.10
d’	5.38	4.88	**8.14**	**9.40**	**8.92**	6.68	6.02	**8.64**	**10.16**	**8.94**	**9.66**

Note: ϵ/C-Value of the sphericity parameter/Pairing of covariance marices and group sizes condition; IGA- Huynh’s (1978) Improved General Approximation Test, WJ-Welch-James test (Keselman et al.,1993), $\bar{\epsilon}$-EBAR (Quintana & Maxwell, 1994),T^2-Hotelling’s test, HL-Hotelling-Lawley Trace, P-Pillai test, W-Wilks test, EB-Empirical Bayes.

Table 3. Empirical Type I Error Rates (%) (**Z** *t*=3 Distributed)

ϵ/C	Main Effect Test					Interaction Effect Test					
	IGA	WJ	ϵ	T^2	EB	IGA	WJ	ϵ	HL(EB)	P(EB)	W(EB)
f=3, *N*=18											
1.0/a	4.78	---	4.68	5.10	4.74	6.70	---	5.72	5.52	4.26	5.12
b	5.58	---	5.66	6.68	6.20	6.72	---	6.62	**7.92**	5.10	6.58
c	4.90	---	2.54	**2.18**	**2.18**	6.50	---	3.52	3.66	**1.86**	2.96
d	6.30	---	**9.40**	**12.88**	**12.48**	**7.80**	---	**11.28**	**16.12**	**12.00**	**14.46**
c’	4.86	---	**1.50**	**1.22**	**0.94**	6.12	---	2.54	**2.16**	**1.06**	**1.68**
d’	**8.40**	**---**	**13.18**	**18.88**	**17.62**	**8.92**	**---**	**14.64**	**20.96**	**16.66**	**19.46**
.40/a	5.56	---	5.58	5.20	5.18	6.28	---	5.20	4.88	3.38	4.32
b	5.94	---	6.28	6.82	6.54	6.12	---	5.96	**7.92**	5.04	7.02
c	4.64	---	**2.30**	**2.14**	**1.76**	6.16	---	3.48	3.36	**1.68**	2.54
d	7.00	---	**10.20**	**13.88**	**13.04**	**8.36**	---	**11.00**	**16.20**	**12.16**	**15.00**
c’	5.34	---	**1.80**	**1.24**	**1.18**	5.34	---	**2.46**	2.60	**1.20**	**2.10**
d’	**9.00**	**---**	**13.00**	**19.04**	**18.32**	**9.06**	**---**	**13.18**	**19.92**	**15.50**	**18.54**
f=6, *N*=45											
1.0/a	5.34	4.96	5.22	5.14	4.92	6.50	5.14	5.06	4.98	4.54	4.76
b	5.62	5.36	5.62	6.12	5.94	5.98	5.06	5.66	6.12	5.24	5.72
c	5.18	4.74	3.98	3.36	3.38	6.06	5.14	4.04	4.50	3.74	4.20
d	5.16	5.02	6.86	**8.40**	**8.02**	6.94	5.76	**8.18**	**9.74**	**8.46**	**9.26**
c’	4.86	4.64	3.22	**3.08**	**2.88**	5.76	5.12	3.48	3.92	3.32	3.66
d’	5.68	5.36	**8.12**	**10.48**	**10.10**	6.14	5.92	**8.68**	**10.54**	**9.32**	**10.00**
.40/a	4.24	4.74	4.24	4.90	4.84	6.82	5.00	4.78	4.48	4.00	4.28
b	5.60	4.64	5.78	5.56	5.20	6.80	5.62	6.50	6.12	5.30	5.78
c	5.00	4.84	3.72	3.56	3.40	5.66	5.36	4.22	4.66	3.88	4.22
d	5.10	4.96	6.44	**8.44**	**8.20**	6.02	5.46	7.12	**8.72**	**7.58**	**8.16**
c’	5.00	4.98	3.30	2.84	2.82	6.44	5.52	4.12	4.06	3.26	3.78
d’	5.54	5.30	**7.96**	**10.34**	**10.12**	6.20	5.88	**8.08**	**10.52**	**9.32**	**9.92**

Note: ϵ/C-Value of the sphericity parameter/Pairing of covariance marices and group sizes condition; IGA-Huynh’s (1978) Improved General Approximation Test, WJ-Welch-James test (Keselman et al.,1993), ϵ-EBAR (Quintana & Maxwell, 1994),T^2-Hotelling’s test, HL-Hotelling-Lawley Trace, P-Pillai test, W-Wilks test, EB-Empirical Bayes.

Table 4. Empirical Type I Error Rates (%) (**Z** Lognormal Distributed)

C	Main Effect Test					Interaction Effect Test					
	IGA	WJ	$\tilde{\varepsilon}$	T^2	EB	IGA	WJ	$\tilde{\varepsilon}$	HL(EB)	P(EB)	W(EB)
Response Variable Normally Distributed											
f=3, *N*=18, ε=1.00											
a	4.90	---	4.68	4.50	4.42	6.22	---	5.10	5.56	3.98	4.64
b	5.38	---	5.48	6.18	5.52	6.72	---	6.44	**7.94**	5.10	6.68
c	5.14	---	3.04	2.70	2.64	5.86	---	3.42	3.54	**1.86**	2.82
d	6.72	---	**10.14**	**14.08**	**13.38**	**7.64**	**---**	**10.58**	**15.76**	**12.16**	**14.46**
c’	5.18	---	**1.40**	**1.22**	**1.08**	6.16	---	**2.46**	**1.80**	**1.08**	**1.44**
d’	**8.96**	---	**14.22**	**19.54**	**19.06**	**9.68**	**---**	**15.12**	**21.68**	**17.52**	**20.60**
f=6, *N*=45, ε=0.40											
a	5.14	4.48	5.12	4.66	4.66	6.38	5.54	4.66	5.14	4.38	4.80
b	4.82	5.08	4.92	5.88	5.48	5.94	4.74	5.50	6.04	4.96	5.58
c	4.34	4.80	3.30	3.70	3.64	6.60	5.40	4.80	4.54	3.60	4.06
d	5.66	5.40	7.28	**8.38**	**8.20**	6.52	6.20	**7.70**	**9.92**	**8.94**	**9.52**
c’	4.68	4.88	2.90	3.04	2.86	5.94	5.06	3.98	3.70	3.00	3.36
d’	5.46	5.38	**8.00**	**10.02**	**9.70**	6.72	6.34	**8.50**	**10.94**	**9.56**	**10.36**
Response Variable Chi-Square (3) Distributed											
f=3, *N*=18, ε=1.00											
a	6.44	---	6.56	**10.24**	**8.84**	3.94	---	3.96	4.62	3.54	4.26
b	6.20	---	6.48	**11.46**	**10.26**	5.50	---	6.28	**8.20**	5.54	7.02
c	6.32	---	3.88	6.18	4.78	4.80	---	3.54	3.64	**2.12**	3.02
d	**8.66**	---	**12.16**	**19.58**	**18.42**	7.10	---	**10.52**	**14.36**	**10.98**	**13.02**
c’	6.32	---	2.56	4.46	3.26	4.98	---	2.50	2.58	**1.60**	**2.28**
d’	**10.96**	---	**15.62**	**27.50**	**22.48**	**9.36**	---	**14.42**	**19.90**	**15.88**	**18.60**
f=6, *N*=45, ε=0.40											
a	5.40	---	5.46	6.78	6.56	5.86	---	4.88	4.18	3.80	4.06
b	5.10	**---**	5.30	**8.26**	**8.10**	5.78	**---**	5.70	6.46	5.68	6.10
c	5.40	**---**	4.40	7.42	7.12	5.94	**---**	4.60	5.78	4.78	5.30
d	5.76	**---**	7.44	**11.22**	**11.00**	6.14	**---**	7.36	**9.32**	**7.62**	**8.56**
c’	5.24	---	3.62	4.94	4.78	5.62	---	3.92	3.84	2.96	3.54
d’	5.32	**---**	**7.84**	**12.88**	**12.62**	6.10	---	**7.96**	**10.76**	**9.64**	**10.42**

Note: C-Pairing of covariance marices and group sizes condition; IGA-Huynh’s (1978) Improved General Approximation Test, WJ-Welch-James test (Keselman et al., 1993), $\tilde{\varepsilon}$-EBAR (Quintana & Maxwell, 1994), T^2-Hotelling’s test, HL-Hotelling-Lawley Trace, P-Pillai test, W-Wilks test, EB-Empirical Bayes.--- - indicates that rates of Type I error were not collected since sample size was smaller than recommended values.

Table 5. Empirical Type I Error Rates (%) (Normally Distributed Data)

	Main Effect Test			Interaction Effect Test			
C	є̄	T^2	EB	є̄	HL(EB)	P(EB)	W(EB)
N=6, *f*=3, є=1.00							
a	5.92	5.48	5.78	6.14	6.46	4.34	6.42
b	6.24	6.20	6.90	6.92	**8.16**	4.58	7.38
N=6, *f*=3, є=0.40							
a	5.54	5.58	6.08	5.60	6.22	2.52	5.62
b	6.30	5.70	7.12	6.84	**7.64**	2.98	6.94
N=6, *f*=6, є=0.75							
a	4.80	5.76	5.88	4.94	6.96	5.54	6.84
b	5.34	6.72	7.28	5.88	**8.92**	6.48	**8.60**
N=9, *f*=3, є=1.00							
a	5.42	4.84	5.48	6.04	6.98	3.68	6.22
b	6.20	7.08	7.32	6.96	8.66	4.24	7.26
N=9, *f*=3, є=0.40							
a	4.96	4.56	5.24	5.26	5.76	2.76	4.70
b	6.22	7.48	**7.66**	6.66	**8.70**	3.50	7.28
N=9, *f*=6, є=0.75							
a	4.70	4.78	5.14	5.52	6.90	4.36	6.10
b	5.58	7.32	6.52	5.56	**8.54**	5.30	**7.66**

Note: C-Pairing of covariance marices and group sizes condition; є̄- EBAR (Quintana &Maxwell, 1994), T^2-Hotelling's test, HL-Hotelling-Lawley Trace, P-Pillai test, W-Wilks test, EB-Empirical Bayes.

Table 6 Empirical Type I Error Rates (%) [Chi-Square (3) Data]

	Main Effect Test			Interaction Effect Test			
C	$\bar{\epsilon}$	T^2	EB	$\bar{\epsilon}$	HL(EB)	P(EB)	W(EB)
N=6, *f*=3, ϵ=1.00							
a	**8.04**	6.44	8.66	5.80	6.66	3.92	6.38
b	**8.76**	**8.06**	**9.68**	6.62	**8.24**	5.16	**7.96**
N=6, *f*=3, ϵ=0.40							
a	7.50	7.06	**8.64**	4.90	5.38	2.30	5.14
b	**8.24**	**7.94**	**10.44**	6.50	**7.72**	3.00	7.10
N=6, *f*=6, ϵ=0.75							
a	6.04	6.58	**7.70**	4.58	6.86	5.24	6.64
b	6.82	**8.46**	**9.22**	5.42	**8.14**	5.60	**7.76**
N=9, *f*=3, ϵ=1.00							
a	7.26	**10.52**	**9.38**	5.18	6.02	3.44	5.24
b	**7.94**	**12.78**	**11.08**	6.58	**8.66**	4.40	7.32
N=9, *f*=3, ϵ=0.40							
a	7.32	**10.34**	**10.34**	4.44	5.80	2.48	4.88
b	**7.66**	**11.94**	**11.10**	6.04	**8.84**	3.68	7.28
N=9, *f*=6, ϵ=0.75							
a	5.66	**9.74**	**8.12**	4.24	5.42	3.58	4.90
b	6.90	**13.02**	**10.30**	5.62	**8.04**	4.56	6.82

Note: C-Pairing of covariance marices and group sizes condition; $\bar{\epsilon}$-EBAR (Quintana &Maxwell, 1994), T^2-Hotelling's test, HL-Hotelling-Lawley Trace, P-Pillai test, W-Wilks test, EB-Empirical Bayes.