

The effect of skewness and kurtosis on the robustness of linear mixed models

Jaume Arnau · Rebecca Bendayan · María J. Blanca · Roser Bono

Published online: 9 January 2013
© Psychonomic Society, Inc. 2013

Abstract This study analyzes the robustness of the linear mixed model (LMM) with the Kenward–Roger (KR) procedure to violations of normality and sphericity when used in split-plot designs with small sample sizes. Specifically, it explores the independent effect of skewness and kurtosis on KR robustness for the values of skewness and kurtosis coefficients that are most frequently found in psychological and educational research data. To this end, a Monte Carlo simulation study was designed, considering a split-plot design with three levels of the between-subjects grouping factor and four levels of the within-subjects factor. Robustness is assessed in terms of the probability of type I error. The results showed that (1) the robustness of the KR procedure does not differ as a function of the violation or satisfaction of the sphericity assumption when small samples are used; (2) the LMM with KR can be a good option for analyzing total sample sizes of 45 or larger when their distributions are normal, slightly or moderately skewed, and with different degrees of kurtosis violation; (3) the effect of skewness on the robustness of the LMM with KR is greater than the corresponding effect of kurtosis for common values; and (4) when data are not normal and the total sample size is 30, the procedure is not

robust. Alternative analyses should be performed when the total sample size is 30.

Keywords Linear mixed model · Kenward–Roger procedure · Skewness · Kurtosis · Robustness

Longitudinal studies, which can be broadly defined as those studies in which the response of each individual is observed on two or more occasions, play a prominent role in the behavioral sciences. The empirical evidence obtained by considering the changes in psychological and educational variables over time can be used to establish predictive relationships that sometimes cannot be detected when cross-sectional studies are used. One of the most popular longitudinal designs is the split-plot design, in which individuals are measured repeatedly on two or more occasions in relation to one or more grouping factors. Data from this design are frequently analyzed with an analysis of variance (ANOVA) with within-subjects and between-subjects factors. This approach is valid under certain assumptions, such as normality, sphericity, and independence of the observations. However, when these assumptions are not satisfied, as is often the case in psychological and educational research (Blanca, Arnau, Bono, López-Montiel & Bendayan, 2012; Jaccard & Ackerman, 1985; Rogan, Keselman & Mendoza, 1979; Winer, 1971), the robustness of ANOVA is not guaranteed (Berkovits, Hancock & Nevitt, 2000; Keselman, Lix & Keselman, 1996), and alternative procedures of data analysis may be necessary.

One of the most suitable approaches for analyzing the data from repeated measures designs, in general, and split-plot designs, in particular, is the linear mixed model (LMM; Cnaan, Laird & Slasor, 1997; Laird & Ware, 1982; Littell, Milliken, Stroup & Wolfinger, 1996). The LMM allows researchers to include random factors and to model the covariance structure of their data prior to testing the

J. Arnau · R. Bono
Department of Methodology of the Behavioral Sciences,
University of Barcelona, Barcelona, Spain

R. Bendayan · M. J. Blanca
Department of Psychobiology and Methodology of Behavioural
Sciences, University of Málaga, Málaga, Spain

R. Bendayan (✉)
Departamento de Psicobiología y Metodología de las Ciencias del
Comportamiento, Facultad de Psicología, Campus de Teatinos,
Universidad de Málaga, 29071 Málaga, Spain
e-mail: bendayan@uma.es

treatment effects. In general, the LMM described by Laird and Ware (1982) can be written as in Eq. 1:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}. \quad (1)$$

where \mathbf{y} is the observations vector, \mathbf{X} the matrix for the fixed effects model, $\boldsymbol{\beta}$ is the vector of the fixed effects parameters, \mathbf{Z} is the matrix for the random effects model, \mathbf{u} the vector of the random effects parameters, and \mathbf{e} is the vector of random errors.

The distribution assumptions of this model are that \mathbf{u} and \mathbf{e} are independent random vectors distributed as $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$, respectively, where \mathbf{G} is a matrix of unknown covariance parameters for the between-subjects random effects and \mathbf{R} is a covariance matrix for the within-subjects errors. Since \mathbf{u} and \mathbf{e} are independent vectors, their covariance is equal to 0, and the covariance matrix of \mathbf{y} is $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$.

The matrices \mathbf{G} and \mathbf{R} are usually unknown, and consequently, an estimate of \mathbf{V} must be used. This is often done by means of the residual maximum likelihood estimation (Zimmerman & Núñez-antón, 2001), as in Eq. 2:

$$\widehat{\mathbf{V}} = \mathbf{V}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \quad (2)$$

Once the covariance matrix has been selected and its parameters estimated, $\boldsymbol{\beta}$ is estimated through the estimated generalized least squares estimator, as in Eq. 3:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{y}. \quad (3)$$

However, the true variance of $\widehat{\boldsymbol{\beta}}$ is not $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ because $\widehat{\boldsymbol{\beta}}$ contains variation due to $\widehat{\mathbf{V}}$, so it is not always a good estimate of \mathbf{V} (Littell, 2002). As Vallejo, Fernández, Herrero and Conejo (2004) highlighted, this means that the likelihood-based inference should be interpreted with caution when the sample size is not large enough.

To summarize, the LMM approach uses statistics that have good large-sample properties but do not appear to be adequate when used with small samples (Wright & Wolfinger, 1996). However, small sample properties can be improved by procedures that adjust the degrees of freedom—for example, the method developed by Kenward and Roger (1997). The KR procedure provides an adjusted estimator of the covariance matrix of $\boldsymbol{\beta}$ that has reduced the bias for small sample inference when the asymptotic covariance matrix underestimates $\widehat{\mathbf{V}}$.

Specifically, the LMM uses Wald-type statistics that can be defined as in Eq. 4:

$$\mathbf{W} = (\mathbf{C}\widehat{\boldsymbol{\beta}})'(\mathbf{C}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}}), \quad (4)$$

where \mathbf{C} is a contrast matrix with range q , and the Wald F for the hypothesis $H_0: \mathbf{C}\boldsymbol{\beta} = 0$ is $F = \mathbf{W}/q$.

If we calculate a scale factor δ and an approximate value for the degrees of freedom ν , then the F statistic for the KR method is given by Eq. 5:

$$F^* = \delta F_{KR} = \frac{\delta}{q} (\mathbf{C}\widehat{\boldsymbol{\beta}})' (\mathbf{C}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{C}')^{-1} (\mathbf{C}\widehat{\boldsymbol{\beta}}). \quad (5)$$

The moments of F^* are generated and matched to the moments of the distribution F so as to solve δ and ν . Under the null hypothesis, it is assumed that F^* is approximately distributed in the same way as F , with q degrees of freedom in the numerator and ν degrees of freedom in the denominator. Hence, two values from the data have to be calculated: the degrees of freedom in the denominator ν and a scale factor δ , following Eqs. 6, 7, and 8. Thus,

$$\nu = 4 + \frac{q+2}{q\nu-1}, \quad (6)$$

where

$$y = \frac{V[F_{KR}]}{2E[F_{KR}]^2} \quad (7)$$

and

$$\delta = \frac{\nu}{E[F_{KR}](\nu-2)}. \quad (8)$$

Several simulation studies have examined the use of the LMM with the KR procedure by exploring robustness in split-plot designs when the assumptions of the LMM are not met. In this context, robustness means that the empirical alpha found in the simulations is close to the nominal value of alpha (probability of type I error). Most simulation studies use Bradley's (1978) liberal criterion, according to which the test procedure is considered robust if the empirical type I error is between .025 and .075, for an alpha level of .05.

With respect to the repeated measures effect, Monte Carlo simulation studies have found that the KR procedure is robust to variance heterogeneity with assumed sphericity and different violations of normality—for example, log-normal (Kowalchuk, Keselman, Algina & Wolfinger, 2004), chi-square with three degrees of freedom (Vallejo et al., 2004), or some unknown distributions with moderate (skewness = 1 and kurtosis = 0.75), high (skewness = 1.75 and kurtosis = 3), or very extreme (skewness = 3 and kurtosis = 21) violation of normality (Livacic-Rojas, Vallejo & Fernández, 2006, 2010). As regards the interaction effect, the results are inconsistent. With sphericity assumed, Kowalchuk et al. (2004) found that the procedure was robust when the distribution was log-normal, while Vallejo et al. (2004) showed that KR is conservative with chi-square distributions with three degrees of freedom. However, with different unknown distributions, such as those cited above, studies have shown that KR may be robust (Livacic-Rojas et

al., 2010), conservative (Livacic-Rojas et al., 2006, 2010), or liberal under some sample size conditions when variances are heterogeneous (Vallejo & Ato, 2006).

Arnau, Bono, Blanca and Bendayan (2012) examined KR robustness when the assumptions of normality, sphericity, and variance homogeneity are not met jointly. Specifically, they explored KR robustness with log-normal, exponential, and double exponential distributions. They found that for the repeated measures and the interaction effects, KR was least robust when the distribution was log-normal, in which case it was nearly always liberal when the sphericity assumption was not met. Furthermore, they suggested that skewness and kurtosis could have a differential effect on KR robustness—namely, that higher values of skewness appeared to be related to greater type I error rates, while higher values of kurtosis seemed to be related to reduced type I error rates. However, the effect of kurtosis and skewness on the KR procedure has yet to be specifically explored. Some studies focusing on other statistical procedures have found that the effects of kurtosis are greater than those of skewness (Harwell, Rubinstein, Hayes & Olds, 1992; Hopkins & Weeks, 1990), whereas others have reported that the effects of skewness are greater than those of kurtosis (Arnau et al., 2012; Chafin & Rhiel, 1993; Scheffé, 1959).

The aim of the present study was to analyze the robustness of the LMM, with the KR procedure, to violations of normality and sphericity when used in split-plot designs with small sample sizes. Specifically, we sought to examine whether skewness and kurtosis have a differential effect on KR robustness by exploring both independently. To this end, a simulation study was designed, including the values of skewness and kurtosis coefficients most frequently found in psychological and educational research data (Blanca et al., 2012), as well as the sample sizes most frequently used (Fernández, Vallejo, Livacic-Rojas & Tuero, 2010; Keselman, Huberty, Lix, Olejnik, Cribbie, Donahue Kowalchuk et al., 1998).

Method

A Monte Carlo simulation study was designed to compare the effects of skewness and kurtosis on KR robustness, the comparison being based on type I error rates. This study considered a split-plot design with three levels of the between-subjects grouping factor and four levels of the within-subjects factor—that is, three groups of a number of individuals who are measured on four occasions.

Normal data were generated using a series of macros created ad hoc in SAS 9.2 (SAS Institute, 2008). First, covariance matrices were generated with sphericity values of .57 and .75. Second, the RANNOR generator was used to obtain normally distributed pseudorandom observations, applying the Cholesky factor of the covariance matrix R . Nonnormal data were generated via the same procedure

but were transformed by means of Fleishman coefficients (Fleishman, 1978) corresponding to each of the distributions studied. The within-subjects, between subjects, and interaction effects were set to zero in the population model.

All data were generated assuming variance homogeneity and using the unstructured (UN) covariance structure, since this is the most common approach in behavioral and educational longitudinal data. Indeed, some studies recommend using this structure when the number of observations is moderate or sample sizes are small (Chen & Wei, 2003; Kowalchuk et al., 2004).

The following variables were examined: (1) total sample size, (2) equal and unequal group size, (3) distributional shape of the response variable, and (4) sphericity. Total sample sizes of $N = 30, 45,$ and 60 were considered. These sample sizes correspond to what is most frequently used in behavioral and educational research (Fernández et al., 2010; Keselman et al., 1998; Livacic-Rojas et al., 2006). For each value of N , both equal and unequal group sizes were considered. Unequal group sizes, in which the number of individuals decreases, were considered because unbalanced data due to experimental mortality is very common in longitudinal studies (Keselman et al., 1998). Specifically, with unequal group size, the coefficient of sample size variation, Δn_j , was .33, while the group sizes were as follows: 14, 10, 6 ($N = 30$); 21, 15, 9 ($N = 45$); and 28, 20, 12 ($N = 60$). When the group sizes were equal, $\Delta n_j = 0$, the group sizes were 10, 10, 10 ($N = 30$); 15, 15, 15 ($N = 45$); and 20, 20, 20 ($N = 60$). The coefficient of sample size variation, Δn_j , can be defined as in Eq. 9:

$$\Delta n_j = \frac{\sqrt{\sum_{j=1}^J (n_j - \bar{n})^2 / J}}{\bar{n}}, \quad (9)$$

where n_j is the sample size of each group, J is the number of groups, and $\bar{n} = \sum n_j / J$.

In order to explore the differential effect of skewness (γ_1) and kurtosis (γ_2) on KR robustness, several distributional shapes of the response variable were considered. In a first step, data were generated to be normally distributed, so as to set a baseline. Different values of the γ_1 and γ_2 coefficients were chosen on the basis of a recent study that assessed the distributional shape of real data by examining the values of γ_1 and γ_2 in small samples of educational and behavioral research data (Blanca et al., 2012). This study revealed that γ_1 usually ranges between -2.49 and 2.33 , while γ_2 usually ranges between -1.92 and 7.41 . The values of the γ_1 and γ_2 coefficients (see Table 1) were chosen according to the cutoff points for the typical degree of contamination found in this type of data, as proposed by Blanca et al. (2012). It should be noted that γ_2 is equal to $\beta_2 - 3$, where β_2 is the Pearson coefficient of kurtosis and 3 is the value of β_2 for the normal distribution.

In order to analyze KR robustness to violations of normality and sphericity together, two indices of sphericity were used. A

Table 1 Values of γ_1 and γ_2 coefficients for the considered distributional shapes of the response variable

	Degree of contamination				
	Slight	Moderate	High	Extreme	Very extreme
Skewness ($\gamma_2 = 0$)	$\gamma_1 = 0.4$	$\gamma_1 = 0.8$	$\gamma_1 = 1.6$	$\gamma_1 = 2$	$\gamma_1 = 2.5$
Kurtosis ($\gamma_1 = 0$)	$\gamma_2 = 0.4$	$\gamma_2 = 0.8$	$\gamma_2 = 1.6$	$\gamma_2 = 2$	$\gamma_2 = 2.5; 3.2; 7.2$

value of $\varepsilon = .75$ was used as a good approximation to sphericity, and a value of $\varepsilon = .57$ was used to represent nonsphericity. Ten thousand replications were performed for each combination at a significance level of .05. This number of replications was chosen in order to ensure reliable results with extremely contaminated distributions (Bendayan, Arnau, Blanca, Bono & Alarcón, 2011; Robey & Barcikowski, 1992).

Results

The empirical type I error rates associated with the repeated measures effect and the interaction effect of the LMM combined with the KR procedure were analyzed for each combination of the study variables. The robustness of this model and procedure was evaluated by means of Bradley's (1978) liberal criterion, according to which a test is robust when the empirical type I error rate is between .025 and .075 for $\alpha = .05$. When the empirical type I error rate is above the upper limit, the test is considered liberal, and when it is below the lower limit, it is considered conservative.

Normally distributed data

Table 2 shows the empirical type I error rates for the repeated measures and interaction effects when the data were normally distributed and the sphericity assumption was not met. The results show that for the repeated measures effect KR was generally robust under the conditions studied. Similar results were obtained for the interaction effect, with one exception: When the total sample size was 30, KR was liberal. No differences were found according to whether or not the group sizes were balanced. The violation of the sphericity assumption had no effect on KR robustness.

Skewed data

Table 3 shows the empirical type I error rates for the repeated measures and interaction effects when data were skewed. For the repeated measures effect, the results indicate that with slight and moderate skewness, KR was robust with total sample sizes of 45 and 60. However, the procedure appeared to be liberal with a total sample size of 30. Furthermore, with high, extreme, and very extreme skewness, KR was liberal for

all the conditions. For the interaction effect, KR was robust under all conditions, except for those with slight and moderate skewness and a total sample size of 30, in which case it was liberal. No differences were found regarding whether group sizes were balanced or not. The violation of the sphericity assumption had no effect on KR robustness.

Data with different degrees of kurtosis

As can be seen in Table 4, for the repeated measures effect, KR was robust independently of the degree of kurtosis or violation of the sphericity assumption. KR was also robust for the interaction effect, although with a total sample size of 30, it was mainly liberal, independently of the degree of kurtosis or violation of the sphericity assumption. No differences were found in terms of whether group sizes were balanced or not.

Discussion

This study has analyzed the robustness of the LMM, with the KR procedure, to violations of normality and

Table 2 Empirical type I error rates for the repeated measures and interaction effects (nominal value .05) with respect to normally distributed data

<i>N</i>	<i>n</i> ₁	<i>n</i> ₂	<i>n</i> ₃	Δn_j	$\varepsilon = .57$	$\varepsilon = .75$
Repeated measures effect						
30	10	10	10	.00	.074	.070
30	14	10	6	.33	.072	.076
45	15	15	15	.00	.066	.060
45	21	15	9	.33	.068	.068
60	20	20	20	.00	.063	.063
60	28	20	12	.33	.061	.066
Interaction effect						
30	10	10	10	.00	.079	.076
30	14	10	6	.33	.078	.079
45	15	15	15	.00	.068	.060
45	21	15	9	.33	.071	.068
60	20	20	20	.00	.065	.065
60	28	20	12	.33	.060	.062

N, total sample size; *n*_{*j*}, group sample size; Δn_j , coefficient of sample size variation; ε , sphericity. In bold: liberal

Table 3 Empirical type I error rates for the repeated measures and interaction effects (nominal value 0.05) with respect to skewed data

Degree of skewness														
N	n ₁	n ₂	n ₃	Δn _j	Slight		Moderate		High		Extreme		Very extreme	
					γ ₁ = 0.4		γ ₁ = 0.8		γ ₁ = 1.6		γ ₁ = 2		γ ₁ = 2.5	
					ε = .57	ε = .75	ε = .57	ε = .75	ε = .57	ε = .75	ε = .57	ε = .75	ε = .57	ε = .75
Repeated measures effect														
30	10	10	10	0.00	.071	.077	.088	.077	.259	.237	.267	.241	.151	.138
45	15	15	15	0.00	.067	.069	.074	.070	.216	.192	.228	.202	.126	.120
60	20	20	20	0.00	.063	.065	.068	.071	.177	.164	.195	.175	.113	.106
30	14	10	6	0.33	.076	.075	.082	.079	.252	.215	.247	.224	.137	.129
45	21	15	9	0.33	.067	.061	.077	.070	.210	.188	.210	.186	.116	.104
60	28	20	12	0.33	.059	.061	.065	.066	.171	.152	.193	.166	.101	.095
Interaction effect														
30	10	10	10	0.00	.080	.079	.077	.076	.050	.052	.054	.054	.069	.067
45	15	15	15	0.00	.068	.067	.065	.068	.045	.048	.046	.049	.058	.054
60	20	20	20	0.00	.063	.064	.060	.063	.042	.049	.039	.041	.053	.053
30	14	10	6	0.33	.084	.075	.082	.078	.061	.058	.064	.062	.067	.072
45	21	15	9	0.33	.068	.069	.068	.065	.054	.049	.053	.056	.064	.064
60	28	20	12	0.33	.063	.061	.060	.069	.048	.048	.047	.048	.049	.053

N, total sample size; n_j, group sample size; Δn_j, coefficient of sample size variation; γ₂, skewness; ε, sphericity. In bold: liberal

sphericity when using split-plot designs with small sample sizes. More specifically, its aim was to explore the independent effect of skewness and kurtosis on KR robustness for the values of skewness and kurtosis coefficients that are

Table 4 Empirical type I error rates for the repeated measures and interaction effects (nominal value 0.05), using data with different kurtosis coefficients

Degree of kurtosis																		
N	n ₁	n ₂	n ₃	Δn _j	Slight		Moderate		High		Extreme		Very extreme					
					γ ₂ = 0.4		γ ₂ = 0.8		γ ₂ = 1.6		γ ₂ = 2		γ ₂ = 2.5		γ ₂ = 3.2		γ ₂ = 7.2	
					ε = .57	ε = .75	ε = .57	ε = .75	ε = .57	ε = .75	ε = .57	ε = .75	ε = .57	ε = .75	ε = .57	ε = .75	ε = .57	ε = .75
Repeated measures effect																		
30	10	10	10	0.00	.072	.071	.068	.073	.069	.069	.069	.067	.069	.070	.072	.066	.064	.067
45	15	15	15	0.00	.066	.067	.063	.068	.064	.065	.059	.062	.059	.063	.062	.059	.058	.058
60	20	20	20	0.00	.065	.062	.063	.059	.061	.057	.062	.060	.059	.058	.061	.062	.053	.061
30	14	10	6	0.33	.070	.069	.072	.070	.068	.074	.074	.068	.068	.070	.068	.072	.070	.065
45	21	15	9	0.33	.061	.063	.066	.061	.053	.071	.064	.065	.062	.058	.062	.064	.063	.063
60	28	20	12	0.33	.059	.060	.062	.060	.064	.066	.058	.063	.058	.060	.054	.059	.060	.061
Interaction effect																		
30	10	10	10	0.00	.080	.073	.079	.079	.075	.077	.076	.076	.072	.075	.072	.074	.072	0.063
45	15	15	15	0.00	.065	.065	.059	.065	.060	.064	.063	.061	.065	.064	.059	.064	.054	0.061
60	20	20	20	0.00	.061	.058	.062	.059	.059	.059	.060	.062	.055	.061	.061	.060	.054	0.061
30	14	10	6	0.33	.074	.081	.076	.079	.076	.079	.078	.077	.077	.074	.077	.076	.074	0.070
45	21	15	9	0.33	.070	.069	.062	.068	.053	.066	.067	.067	.066	.066	.066	.062	.064	0.065
60	28	20	12	0.33	.057	.065	.061	.065	.060	.059	.059	.061	.059	.061	.058	.060	.058	0.058

N, total sample size; n_j, group sample size; Δn_j, coefficient of sample size variation; γ₁, kurtosis; ε, sphericity. In bold: liberal

most frequently found in psychological and educational research data.

The results showed that for the repeated measures effect the LMM with KR was robust mainly when data were normal, regardless of whether the sphericity assumption was met. Likewise, for the interaction effect, the procedure was also robust when the total sample size was 45 or larger, but it was liberal when the total sample size was 30. These results are consistent with the findings of Livacic-Rojas et al. (2006, 2010) and Arnau, Bono and Vallejo (2009), who highlighted the problems with the KR procedure when using very small samples with normal and spherical data.

With regard to the robustness of the procedure when data were slightly or moderately skewed, KR was robust for both the repeated measures and interaction effects with total sample sizes of 45 and 60, but it was liberal with a total sample size of 30 (regardless of whether the sphericity assumption was violated). When data were highly, extremely, or very extremely skewed, the procedure was liberal for the repeated measures effect for all the sample sizes considered, regardless of whether the sphericity assumption was violated. By contrast, KR was robust under all the conditions for the interaction effect when data were highly, extremely, or very extremely skewed. These findings are partially consistent with those reported in studies about other statistical procedures (Chafin & Rhiel, 1993; Scheffé, 1959). Furthermore, as Arnau et al. (2012) pointed out, with small samples, the robustness of the LMM with the KR procedure decreases as skewness increases.

Having explored the effect of skewness on the robustness of the KR procedure, the third phase of the study examined the effect of kurtosis. Here, the results indicated that for the repeated measures effect, the procedure was robust, independently of the degree of kurtosis or violation of the sphericity assumption. KR was also robust for the interaction effect, although with a total sample size of 30, it was mainly liberal, but again independently of the degree of kurtosis or violation of the sphericity assumption. These findings are partially consistent with research that has reported the effect of kurtosis on the robustness of other statistical tests (Harwell et al., 1992; Hopkins & Weeks, 1990). Although the present results show an effect of kurtosis on empirical type I error rates with small samples and in relation to the interaction effect, taken together they support previous studies (Arnau et al., 2012; Chafin & Rhiel, 1993; Scheffé, 1959) that have suggested that skewness effects are greater than those of kurtosis. Specifically, the present results suggest that the effect of skewness is greater for the repeated measures effect, whereas the effect of kurtosis is slightly greater for the interaction effect only when the sample size is 30. With respect to the main aim of this study, the results as a whole indicate that there is an independent effect of skewness and kurtosis on KR robustness. Further research should now examine the effect of both skewness and kurtosis, jointly, on KR robustness, because in psychological

and educational data, measures may be skewed and kurtotic at the same time.

Furthermore, and as described above, the coefficient of sample size variation was varied in the simulation; that is, equal and unequal group sizes were considered. No differences were found in any of the studied conditions in relation to whether or not the group sizes were balanced. The results suggest that KR robustness could be more affected by the total sample size when small samples are used and the skewness of the response variable, rather than by whether the design is balanced or not. In this context, it would be interesting for future research to explore the effect of different unequal group sizes by considering other coefficients of sample size variation.

It should be noted that comparison of the present results with previous findings may be hampered by the use of different distributions. Indeed, our results are limited to the range of conditions examined, although they may nonetheless help to decide whether the LMM is suitable for use with specific nonnormal data. As a final note, the following conclusions can be drawn regarding use of the LMM with the KR procedure. First, the robustness of this procedure does not differ as a function of the violation or satisfaction of the sphericity assumption when small samples are used. Second, the LMM with KR can be a good option for analyzing total sample sizes of 45 or larger when their distributions are normal, slightly skewed, or moderately skewed, and with different degrees of kurtosis. Third, for the repeated measures effect, the effect of skewness on the robustness of LMM with KR is greater than the corresponding effect of kurtosis. Finally, when data are not normal and the total sample size is 30, the procedure is not robust, and alternative analyses should be performed.

Considering the results obtained, as well as the fact that small sample sizes and nonnormal data are frequent in longitudinal psychological and educational research (Blanca et al., 2012; Fernández et al., 2010; Keselman et al., 1998; Lei & Lomax, 2005; Micceri, 1989), further studies are now required to explore the robustness of the LMM with other nonnormal unknown distributions, different total and group sample sizes, and a greater number of observations.

Acknowledgments This study was supported by grant PSI2009-11136 from the Spanish Ministry of Science and Innovation.

References

- Arnau, J., Bono, R., Blanca, M. J., & Bendayan, R. (2012). Using the linear mixed model to analyze non-normal data distributions in longitudinal designs. *Behavior Research Methods*. doi:10.3758/s13428-012-0196-y
- Arnau, J., Bono, R., & Vallejo, G. (2009). Analyzing small samples of repeated measures data with the mixed-model adjusted F test. *Communications in Statistics. Simulation and Computations*, 38, 1083–1103. doi:10.1080/03610910902785746

- Bendayan, R., Arnau, J., Blanca, M. J., Bono, R., & Alarcón, R. (2011). Dos procedimientos para generar datos no normales (Fleishman vs Ramberg) y cantidad de simulaciones. Estudio comparativo. *XII Congreso de Metodología de las Ciencias del Comportamiento*. San Sebastian.
- Berkovits, I., Hancock, G. R., & Nevitt, J. (2000). Bootstrap resampling approaches for repeated measure designs: Relative robustness to sphericity and normality violations. *Educational and Psychological Measurement*, 60(6), 877–892. doi:10.1177/00131640021970961
- Blanca, M. J., Arnau, J., Bono, R., López-Montiel, D., & Bendayan, R. (2012). Skewness and kurtosis in real data samples. *Methodology. European Journal of Research Methods for the Behavioral and Social Sciences*. doi:10.1027/1614-2241/a000057
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144–152. doi:10.1111/j.2044-8317.1978.tb00581.x
- Chafin, W. W., & Rhiel, S. G. (1993). The effect of skewness and kurtosis on the one-sample *T* test and the impact of knowledge of the population standard deviation. *Journal of Statistical Computation and Simulation*, 46(1–2), 79–90. doi:10.1080/00949659308811494
- Chen, X., & Wei, L. (2003). A comparison of recent methods for the analysis of small-sample cross-over studies. *Statistics in Medicine*, 22, 2821–2833. doi:10.1002/sim.1537
- Cnaan, A., Laird, N. M., & Slasor, P. (1997). Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, 16, 2349–2380. doi:10.1002/(SICI)1097-0258(19971030)16:20<2349::AID-SIM667>3.0.CO;2-E
- Fernández, P., Vallejo, G., Livacic-Rojas, P., & Tuero, E. (2010). Características y análisis de los diseños de medidas repetidas en la investigación experimental en España en los últimos 10 años. *Actas del XI Congreso de Metodología de las Ciencias del Comportamiento*. Málaga.
- Fleishman, A. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43(4), 521–531. doi:10.1007/BF02293811
- Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one and two factors fixed effect ANOVA case. *Journal of Educational Statistics*, 17, 315–339. doi:10.2307/1165127
- Hopkins, K. D., & Weeks, D. L. (1990). Test for normality and measures of skewness and kurtosis: Their place in research reporting. *Educational and Psychological Measurement*, 50, 717–729. doi:10.1177/0013164490504001
- Jaccard, J., & Ackerman, L. (1985). Repeated measures analysis of means in clinical research. *Journal of Consulting and Clinical Psychology*, 53, 426–428. doi:10.1037/0022-006X.53.3.426
- Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53, 983–997. doi:10.2307/2533558
- 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 education researchers: An analysis of the ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68, 350–386.
- Keselman, J. C., Lix, L. M., & Keselman, H. J. (1996). The analysis of repeated measurements: A quantitative research synthesis. *British Journal of Mathematical and Statistical Psychology*, 49, 275–298. doi:10.1111/j.2044-8317.1996.tb01089.x
- Kowalchuk, R. K., Keselman, H. J., Algina, J., & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted *F* tests. *Educational and Psychological Measurement*, 64(2), 224–242. doi:10.1177/0013164403260196
- Laird, N. M., & Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, 38, 963–974. doi:10.2307/2529876
- Lei, M., & Lomax, R. G. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling*, 12, 1–27. doi:10.1207/s15328007sem1201_1
- Littell, R. C. (2002). Analysis of unbalanced mixed models data: A case study for comparison of ANOVA versus REM/GLS. *Journal of Agricultural, Biological, and Environmental Statistics*, 7, 472–490. doi:10.1198/108571102816
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS System for mixed models*. Cary: SAS Institute Inc.
- Livacic-Rojas, P., Vallejo, G., & Fernández, P. (2006). Procedimientos estadísticos alternativos para evaluar la robustez mediante diseños de medidas repetidas. *Revista Latinoamericana de Psicología*, 38(3), 579–598.
- Livacic-Rojas, P., Vallejo, G., & Fernández, P. (2010). Analysis of Type I error rate of univariate and multivariate procedures in repeated measures designs. *Communications in Statistics - Simulation and Computation*, 39, 624–640. doi:10.1080/03610910903548952
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156–166. doi:10.1037/0033-2909.105.1.156
- Rogan, J. C., Keselman, H. J., & Mendoza, J. L. (1979). Analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology*, 32, 269–286. doi:10.1111/j.2044-8317.1979.tb00598.x
- Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, 45, 283–288. doi:10.1111/j.2044-8317.1992.tb00993.x
- SAS Institute Inc. (2008). *SAS/STAT 9.2 user's guide*. Cary, NC: SAS Institute Inc.
- Scheffé, H. (1959). *The analyses of variance*. New York: Wiley.
- Vallejo, G., & Ato, M. (2006). Modified Brown–Forsythe procedure for testing interaction effects in split-plot designs. *Multivariate Behavioral Research*, 41, 549–578. doi:10.1207/s15327906mbr4104_6
- Vallejo, G., Fernández, P., Herrero, J., & Conejo, N. M. (2004). Alternative procedures for testing fixed effects in repeated measures designs when assumptions are violated. *Psicothema*, 16(3), 498–508.
- Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.
- Wright, S. P., & Wolfinger, R. D. (1996). Repeated measures analysis using mixed models: Some simulation results. *Conference on Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions*. Nantucket, MA.
- Zimmerman, D. L., & Núñez-antón, V. (2001). Parametric modeling of growth curve data: An overview. *TEST*, 10, 111–186. doi:10.1007/BF02595823