GRUPO EDITORIAL DE PSICOFUNDACIÓN

Psicothema (2025) 37(3) 12-22

Psicothema

https://www.psicothema.com/es • ISSN 0214-9915 • eISSN 1886-144X

Colegio Oficial de Psicología del Principado de Asturias

Article

When to use Bootstrap-*F* in One-Way Repeated Measures ANOVA: Type I Error and Power

María J. Blanca[1], Roser Bono[2,3], Jaume Arnau[3], F. Javier García-Castro[4], Rafael Alarcón[1] and Guillermo Vallejo[5]

[1]University of Malaga (Spain)
[2]Institute of Neurosciences, University of Barcelona (Spain)
[3]University of Barcelona (Spain)
[4]Universidad Loyola Andalucía (Spain)
[5]University of Oviedo (Spain)

ARTICLE INFO

Received: 18/12/2024
Accepted: 24/02/2025

Keywords:
Bootstrap-*F*
Within-subject design
Greenhouse-Geisser adjustment
Huynh-Feldt adjustment
Robustness

ABSTRACT

Background: With repeated measures, the traditional ANOVA *F*-statistic requires fulfillment of normality and sphericity. Bootstrap-*F (B-F)* has been proposed as a procedure for dealing with violation of these assumptions when conducting a one-way repeated measures ANOVA. However, evidence regarding its robustness and power is limited. Our aim is to extend knowledge about the behavior of *B-F* with a wider range of conditions. **Method**: A simulation study was performed, manipulating the number of repeated measures, sample sizes, epsilon values, and distribution shape. **Results**: *B-F* may become conservative with higher values of epsilon, and liberal under extreme violation of both normality and sphericity and small sample sizes. In these cases, *B-F* may be used with a more stringent alpha level (.025). The results also show that power is affected by sphericity: the lower the epsilon value, the larger the sample size required to ensure adequate power. **Conclusions**: *B-F* is robust under non-normality and non-sphericity with sample sizes larger than 20-25.

Cuándo Usar *F*-Bootstrap en ANOVA Unifactorial de Medidas Repetidas: Error de Tipo I y Potencia

Palabras clave:
Remuestreo
Diseño intrasujeto
Ajuste Greenhouse-Geisser
Ajuste Huynh-Feldt
Robustez

RESUMEN

Antecedentes: El estadístico *F* del ANOVA de medidas repetidas requiere el cumplimiento de los supuestos de normalidad y esfericidad. El procedimiento *F*-bootstrap (*F-B*) se ha propuesto como alternativa al ANOVA cuando se violan estos supuestos. Sin embargo, la evidencia empírica sobre su robustez y potencia es limitada. El objetivo es analizar el comportamiento de *F-B* en un mayor número de condiciones. **Método:** Se realizó un estudio de simulación, manipulando el número de medidas repetidas, tamaño muestral, valores de épsilon y forma de la distribución. **Resultados**: El procedimiento *F-B* resulta conservador con valores altos de épsilon, y puede llegar a ser liberal bajo una violación extrema de la normalidad y esfericidad con tamaño muestral pequeño. En estos casos, *F-B* puede utilizarse con un nivel de alfa más restrictivo (.025). Los resultados también muestran que la potencia se ve afectada por la esfericidad: cuanto menor es el valor de épsilon, mayor es el tamaño muestral necesario para garantizar una potencia adecuada. **Conclusiones:** El procedimiento *F-B* es robusto en condiciones de no normalidad y no esfericidad con tamaños de muestra superiores a 20-25.

Cite as: Blanca, M. J., Bono, R., Arnau, J., García-Castro, F. J., Alarcón, R., & Vallejo, G. (2025). When to use bootstrap-*F* in one-way repeated measures ANOVA: Type I error and power. *Psicothema, 37*(3), 12-22. https://doi.org/10.70478/psicothema.2025.37.20

Corresponding author: María J. Blanca, blamen@uma.es

Introduction

Bootstrapping is a computing-intensive method introduced by Efron (1979) and colleagues (e.g., Efron & Gong, 1983; Efron & Tibshirani, 1993) that basically involves drawing random samples from the original dataset with replacement, and then computing the sample distribution for a given statistic for each bootstrap sample. This resampling process enables the estimation of confidence intervals, standard errors, and hypothesis tests, providing a robust alternative to traditional parametric methods. The method has a wide range of applications, including comparison of means tests, correlation and regression, multilevel analysis, mediation and moderation, graph analysis, time series analysis, and survival analysis (Chernick & LaBudde, 2011; Christensen & Golino, 2021; Hayes, 2017; Vallejo et al., 2013; Wilcox, 2022). The increasing popularity of bootstrapping for statistical inference has seen it gradually incorporated into the most common statistical software, such as R, SAS and IBM SPSS.

Bootstrap can be used in conjunction with different statistical procedures, including those derived from the general linear model such as regression analysis and analysis of variance (ANOVA), to make inferences about a population. In the case of ANOVA, this involves generating the empirical sampling distribution for the *F*-statistic by repeatedly resampling with replacement from the dataset, rather than using the theoretical distribution of the statistic. Because bootstrap does not rely on the parametric assumptions of normality and homogeneity of variance, it is particularly useful when these assumptions are violated (Chernick, 2008).

Simulation studies are valuable tools that involve running numerous random data sets to assess how a statistic performs under various conditions. Robustness in terms of Type I error is typically interpreted using Bradley's liberal criterion (1978), which considers a statistic to be robust if its Type I error rate is between 2.5% and 7.5% for an alpha of 5%.

When repeated measures are involved, traditional ANOVA (RM-ANOVA) requires normality and sphericity. Simulation studies have shown that the *F*-statistic of RM-ANOVA is generally robust to non-normality when the sphericity assumption is met (Berkovits et al., 2000; Blanca et al., 2023a; Keselman et al., 1996; Kherad-Pajouh & Renaud, 2015). Blanca et al. (2023a) found that the test was robust in 99.95% of the 1786 conditions studied, and also that the Type I error rate was only greater than .075 (specifically, .078) in the case of a design with four repeated measures, extreme departure from normality (skewness $\gamma_1 = 2.31$, kurtosis $\gamma_2 = 8$), and $N = 10$. However, RM-ANOVA is very sensitive to sphericity violation, rendering it a liberal test (Berkovits et al., 2000; Blanca et al., 2023b; Haverkamp & Beauducel, 2017, 2019; Voelkle & McKnight, 2012).

To control Type I error when sphericity is violated, the use of adjusted *F*-tests, such as the Greenhouse-Geisser (*F-GG*) and Huynh-Feldt (*F-HF*) adjustments, has been proposed. These two procedures modify the degrees of freedom of the *F*-statistic by a multiplicative factor, known as epsilon (ε), making it a more demanding test. The value of ε is considered an indicator of the amount by which the data depart from sphericity, and it ranges between $1/k$-1 and 1, where k is the number of repeated measures. When the data meet the sphericity assumption, $\varepsilon = 1$, and the greater the departure from this value the greater the violation of sphericity. *F-GG* and *F-HF* differ in how ε is computed, and the decision over which procedure to use is controversial. Indeed, there is evidence for the superiority of both *F-GG* (Voelkle & MacKnight, 2012) and *F-HF* (Haverkamp & Beauducel, 2017, 2019; Oberfeld & Franke, 2013), while some studies have found that both offer reasonable control of Type I error (Berkovits et al., 2000; Muller et al., 2007). A value-based strategy has also been proposed based on the expected value of ε. For example, Huynh and Feldt (1976) recommend using *F-GG* if ε is less than .75, and *F-HF* for ε greater than .75. More recently, Blanca et al. (2023b) established another cut-off point based on the results of a simulation study with normal data and a larger number of manipulated conditions than were considered in the aforementioned studies, taking the Greenhouse-Geisser ε estimation ($\hat{\varepsilon}$) as reference. They suggested, as a general rule, using *F-GG* because it is more conservative, although in the event of discrepant results from the two procedures, they recommend using *F-GG* for $\hat{\varepsilon}$ values below .60, and *F-HF* for $\hat{\varepsilon}$ values of .60 or higher.

When normality and sphericity are simultaneously violated, the behavior of adjusted *F*-tests depends on several factors, namely sample size and the degree of violation of both sphericity and normality. Blanca et al. (2024) found that although the aforementioned rule generally holds under non-normality and non-sphericity, there are two exceptions in which neither *F-GG* nor *F-HF* is robust: a) With $N \leq 10$, $\hat{\varepsilon} \leq .60$, and severe deviation from normality ($\gamma_1 = 1.41$, $\gamma_2 = 3$) and, b) with $N \leq 30$, $\hat{\varepsilon} \leq .60$, and extreme deviation from normality ($\gamma_1 = 2$, $\gamma_2 = 6$ and 8). These authors discuss several available analytic alternatives, none of which is free of criticism, highlighting that bootstrapping may be the most promising alternative according to results obtained in other studies (e.g., Berkovits et al., 2000).

Berkovits et al. (2000) proposed a bootstrap method for one-way repeated measures ANOVA, referred to as bootstrap-*F (B-F)*, which generates the bootstrap sample from centered data. They conducted a simulation study to analyze the behavior of this procedure in terms of Type I error with a four repeated measures design, introducing different values of sample size (10, 15, 30, and 60) and epsilon (.48, .57, .75, and 1). Distribution shape was also manipulated so as to include both normal data and distributions labeled as showing slight ($\gamma_1 = 1$, $\gamma_2 = 0.75$), moderate ($\gamma_1 = 1.75$, $\gamma_2 = 3.75$), and severe ($\gamma_1 = 3$, $\gamma_2 = 21$) deviation from normality. The results showed that *B-F* was a robust alternative under violations of sphericity and normality, even in small samples and with severe non-normality, with Type I error rates below 7.5% in all conditions manipulated. However, the test became conservative in some cases with $\varepsilon = 1$.

To our knowledge, the behavior of the *B-F* test proposed by Berkovits et al. (2000) has scarcely been investigated with one-way designs, although it has been studied with split-plot designs. Vallejo et al. (2006) performed a simulation study in which they tested this procedure with a 3x4 design with $N = 30$, 45, and 60, $\varepsilon = .50$, .75, and 1, and the same non-normal distributions as Berkovits et al. (2000). The findings were consistent with those of Berkovits et al. (2000), insofar as the test was robust under non-sphericity and non-normality but tended to be conservative with high values of ε. These results were subsequently confirmed by Vallejo et al. (2010) using a 3x4 design with $N = 30$ and 45, and $\varepsilon = .50$, in which they found that *B-F* controlled Type I error with different non-normal distributions.

Overall, the empirical evidence suggests that *B-F* is a robust procedure for dealing with violations of normality and sphericity. However, this evidence is limited as published simulation studies include a small number of manipulated conditions in terms of repeated measures, sample sizes, sphericity, and distribution shapes. The aim of the present study is therefore to extend knowledge about the robustness and power of *B-F* by considering a wider range of conditions. To this end, we included designs with 3, 4, and 6 repeated measures, sample sizes from 10 to 180, $\hat{\varepsilon}$ values from the corresponding lower bound to 1, and six distributions representing slight to extreme deviations from normality.

Bootstrap-*F*

The goal in using this procedure is to estimate an appropriate critical value when the null hypothesis is true. This is done by centering the data in each repeated measure condition, randomly generating *B* bootstrap samples with replacement from the centered data in each condition, computing the statistics for each bootstrap sample generated, and obtaining an estimate of the distribution of the statistic (Wilcox, 2003, p. 379). Berkovits et al. (2000) consider that the *B-F* procedure comprises the following steps:

1. Organize data in a matrix of *N* participants x *K* measurement occasions. To test the null hypothesis of equality of means among repeated measures, compute the *F*-statistic based on original data, labeled as observed F_o.

$$\begin{bmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & \ddots & \vdots \\ X_{N1} & \cdots & X_{Nk} \end{bmatrix}$$

   The data with 3 repeated measures shown in Table 1 provide an illustration of the procedure. The observed F_o is 92.19.
2. Center the data with the aim of estimating an appropriate critical value of the *F*-statistics, subtracting the respective mean of the kth level of the repeated measure from each observation: $C_{ij} = X_{ik} - \overline{X}_{.k}$. This matrix will have the same distributional properties and the same covariance matrix as the original data (Berkovits et al., 2000). The data matrix is now:

$$\begin{bmatrix} C_{11} & \cdots & C_{1k} \\ \vdots & \ddots & \vdots \\ C_{N1} & \cdots & C_{Nk} \end{bmatrix}$$

   Table 2 displays the data matrix with centered data (for the example shown in Table 1).
3. With the centered data, generate *B* bootstrap samples with replacement by randomly sampling *N* rows of data.

$$\begin{bmatrix} C^*_{11} & \cdots & C^*_{1k} \\ \vdots & \ddots & \vdots \\ C^*_{N1} & \cdots & C^*_{Nk} \end{bmatrix}$$

   In the example, we would generate 599 bootstrap samples, although for illustrative purposes, only 2 are displayed in Table 3.

Table 1
Data Matrix for Illustrative Purposes

ID	Measure 1	Measure 2	Measure 3
1	13	6	4
2	9	7	5
3	10	4	3
4	10	7	4
5	11	6	2
6	10	6	5
7	8	6	5
8	11	8	5
9	12	7	6
10	14	6	4
11	11	5	4
12	12	7	5
M	10.92	6.25	4.33
SD	1.67	1.05	1.07

Table 2
Data Matrix With Centered Data (for the Example Shown in Table 1)

ID	Centered 1	Centered 2	Centered 3
1	2.08	-0.25	-0.33
2	-1.92	0.75	0.67
3	-0.92	-2.25	-1.33
4	-0.92	0.75	-0.33
5	0.08	-0.25	-2.33
6	-0.92	-0.25	0.67
7	-2.92	-0.25	0.67
8	0.08	1.75	0.67
9	1.08	0.75	1.67
10	3.08	-0.25	-0.33
11	0.08	-1.25	-0.33
12	1.08	0.75	0.67

Table 3
Bootstrap Samples 1 and 2 With Centered Data (C)

Bootstrap sample 1				Bootstrap sample 2			
ID	C 1	C 2	C 3	ID	C 1	C 2	C3
1	2.08	-0.25	-0.33	2	-1.92	0.75	0.67
6	-0.92	-0.25	0.67	7	-2.92	-0.25	0.67
10	3.08	-0.25	-0.33	4	-0.92	0.75	-0.33
5	0.08	-0.25	-2.33	9	1.08	0.75	1.67
3	-0.92	-2.25	-1.33	8	0.08	1.75	0.67
11	0.08	-1.25	-0.33	7	-2.92	-0.25	0.67
12	1.08	0.75	0.67	5	0.08	-0.25	-2.33
8	0.08	1.75	0.67	6	-0.92	-0.25	0.67
3	-0.92	-2.25	-1.33	5	0.08	-0.25	-2.33
11	0.08	-1.25	-0.33	2	-1.92	0.75	0.67
10	3.08	-0.25	-0.33	12	1.08	0.75	0.67
8	0.08	1.75	0.67	3	-0.92	-2.25	-1.33
$F_1^* = 3.12$				$F_2^* = 2.66$			

4. Compute *F*-statistics with data from each bootstrap sample, labeled as F_1^*, …, F_B^*, thus creating the empirical sampling distribution of the F-statistic. The *F*-statistics for bootstrap samples 1 and 2 are equal to $F_1^* = 3.12$ and $F_2^* = 2.66$, respectively. Sort the F^* values in ascending order. Suppose that we obtain a set of F^* values after performing 599 bootstrap samples. We then sort these values in ascending order, resulting in the following ranking: (1) $F_3^* = 0.95$, (2) $F_2^* = 2.66$, (3) $F_1^* = 3.12$, …, (569) $F_{328}^* = 5.03$, ..., (599) $F_{430}^* = 52.34$.
5. Estimate the critical value F_c^*, where $c = (1 - \alpha)\ B$. The F_o of step 1 is compared with this critical value, and hence the null hypothesis is rejected if $F_o \geq F_c^*$. For instance, with $\alpha = .05$ and $B = 599$ bootstrap samples, $c = .95 * 599 = 569.05$. The F^* in position 569 will thus be the critical value F_c^*. The proportion of F^* values that are larger than the observed F_o represents the bootstrap *p*-value (Berkovits et al., 2000; Vallejo et al., 2010). The null hypothesis of equality of means is rejected if this *p*-value is less than or equal to .05. In the example, the F^*-statistic in position 569 is the critical value: $F_c^* = 5.03$. As $F_o = 92.19$ is larger than 5.03, the null hypothesis of equality of means among repeated measures is rejected. There is no F^* value larger than F_o, yielding a $p < .001$.

The procedure can be performed using the WRS2 library of R (Mair & Wilcox, 2020), with the rmanovab function and without using trimmed means.

Method

Instrument

A simulation study was carried out using the interactive matrix language (IML) module of SAS 9.4. A series of macros was designed to generate data. Non-normal data were generated using the procedure proposed by Fleishman (1978), which applies a polynomial transformation that simulates data with specific values of skewness and kurtosis. To simulate data with different degrees of sphericity violation, we generated a series of unstructured covariance matrices with different values of $\hat{\varepsilon}$ for each repeated measure condition. The unstructured matrix was used because it is the most general covariance structure (Kowalchuk et al., 2004) and is typically found in longitudinal behavioral data (Arnau et al., 2014; Bono et al., 2010). The probability of the values associated with *B-F* was calculated using PROC GLM of SAS (more details about the simulation procedure with SAS are available upon request from the corresponding author). Five thousand replications were performed for each condition manipulated with $B = 599$ bootstraps, as used elsewhere (Vallejo et al., 2006, 2010). This number was selected based on the recommendation that α should be a multiple of $(B + 1)^{-1}$ (Wilcox, 2022). In addition, simulation studies suggest that in terms of probability coverage, there is little or no advantage to using $B > 599$ when $\alpha = .05$ (Wilcox, 2022).

Procedure

Type I error rates were recorded, reflecting the percentage of false rejections of the null hypothesis at the 5% significance level. Robustness of *B-F* was assessed based on Bradley's (1978) liberal criterion, which considers a procedure to be robust if the Type I error rate is between 2.5% and 7.5% for a nominal alpha of 5%. The procedure is considered conservative if the Type I error rate is below the lower bound, and liberal if it is above the upper bound. This criterion was chosen because it is widely used in simulation studies and in research focused on repeated measures (e.g., Arnau et al., 2012; Berkovits et al., 2000; Keselman et al., 1999; Kowalchuk et al., 2004; Livacic-Rojas et al., 2010; Oberfeld & Franke, 2013; Vallejo et al., 2006, 2010, 2011), thus facilitating the comparison of results across similar studies.

The variables manipulated for a one-way design were:

1. Number of repeated measures (*K*): The repeated measures were 3, 4, and 6.
2. Total sample size (*N*): Sample sizes were 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100, 120, 150, and 180.
3. Epsilon ($\hat{\varepsilon}$): The Greenhouse-Geisser estimation of epsilon was used (Box, 1954; Geisser & Greenhouse, 1958; Greenhouse & Geisser, 1959). Depending on the number of repeated measures, $\hat{\varepsilon}$ values ranged from the lower limit to 1. For $K = 3$, $\hat{\varepsilon}$ values were .50, .60, .70, .80, .90, and 1; for $K = 4$, they were .33, .40, .50, .60, .70, .80, .90, and 1; and for $K = 6$, they were .20, .30, .40, .50, .60, .70, .80, .90, and 1.
4. Distribution shape: Six distributions were used, representing slight to extreme deviations from normality, chosen from among those used by Blanca et al. (2024). Skewness and kurtosis values are shown in Table 4.

Empirical power was also calculated as the percentage rejection of the null hypothesis at a significance level of 5%. It was analyzed by selecting mean values with a linear pattern in which the means increase linearly and proportionally to each other (e.g., 0, 0.5, 1), with medium effect size, $f \approx 0.25$. The number of repeated measures and *N* were the same as those for Type I error. Epsilon values ($\hat{\varepsilon}$) ranged from the lower limit to .90. To simplify the study, distributions 2, 3, and 6 were selected so as to represent the variability of performance of *B-F* with respect to Type I error (i.e., *B-F* performed similarly in distributions 3 and 4, and also in distributions 5 and 6). These distributions correspond to moderate ($\gamma_1 = 1$, $\gamma_2 = 1.50$), severe ($\gamma_1 = 1.41$, $\gamma_2 = 3$), and extreme deviation from normality ($\gamma_1 = 2.31$, $\gamma_2 = 8$).

Table 4

Skewness (γ_1) and Kurtosis (γ_2) Coefficients for Each Simulated Distribution

Distribution	Type	γ_1	γ_2
1	-	0.4	0.8
2	Gamma (α = 4)	1	1.50
3	Gamma (α = 2)	1.41	3
4	Gamma (α = 1.5)	1.63	4
5	Exponential	2	6
6	Gamma (α = 0.75)	2.31	8

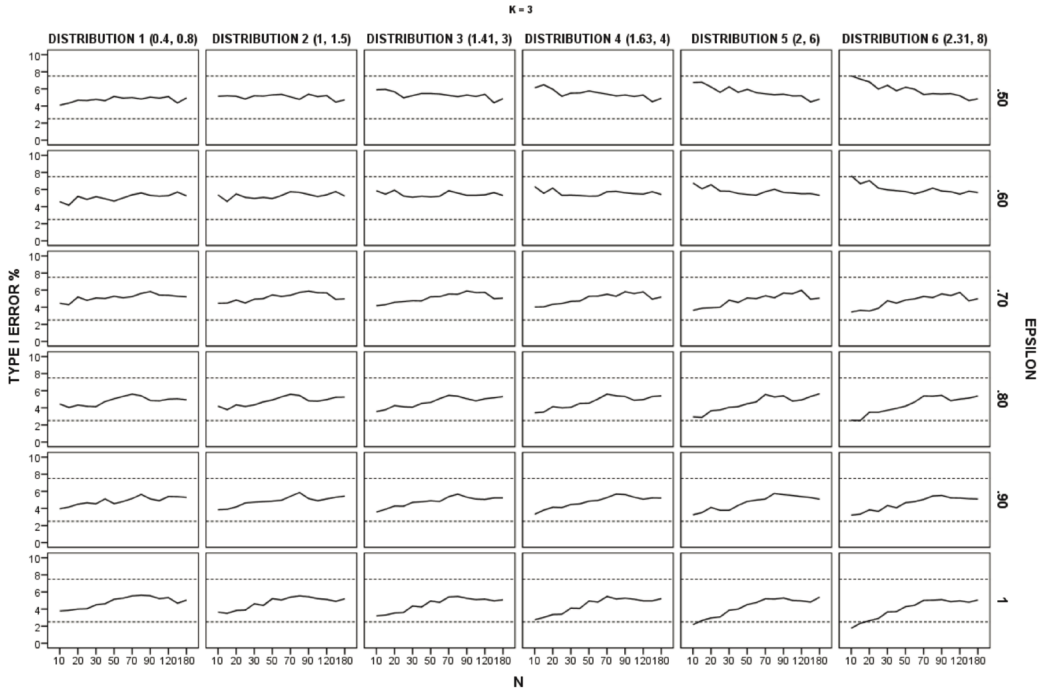
Results

Type I Error Rate

Type I error rates for each *K*, distribution, *N*, and $\hat{\varepsilon}$ value are displayed in Figures 1-3 (detailed tables can be found at https://dx.doi.org/10.24310/riuma.37706). The results are summarized in Table 5**.** For $K = 3$ and 4, *B-F* is robust with distributions 1-4, with maximum values of γ_1 and γ_2 equal to 1.63 and 4, respectively. For the distribution with $\gamma_1 = 2$ and $\gamma_2 = 6$, the procedure is conservative

Figure 1

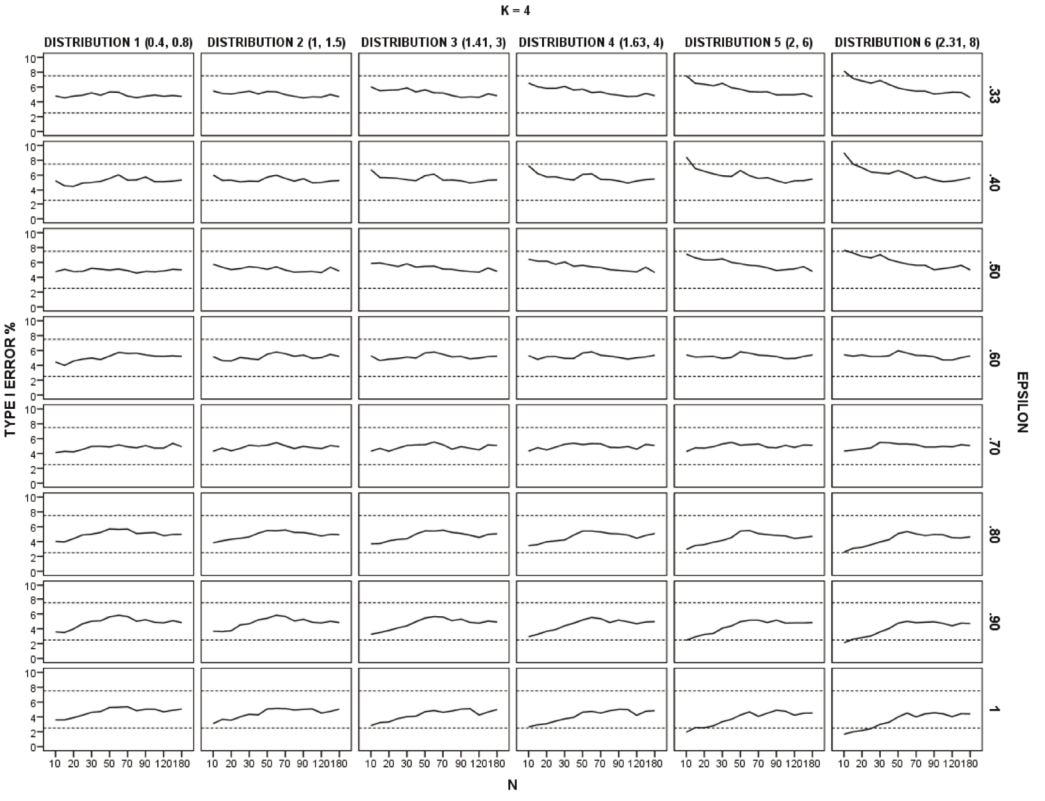
Type I Error Rate (Percentage) for K = 3 as a Function of Distribution Shape, N, and ε̂



Note. In parentheses: skewness and kurtosis coefficients.

Figure 2

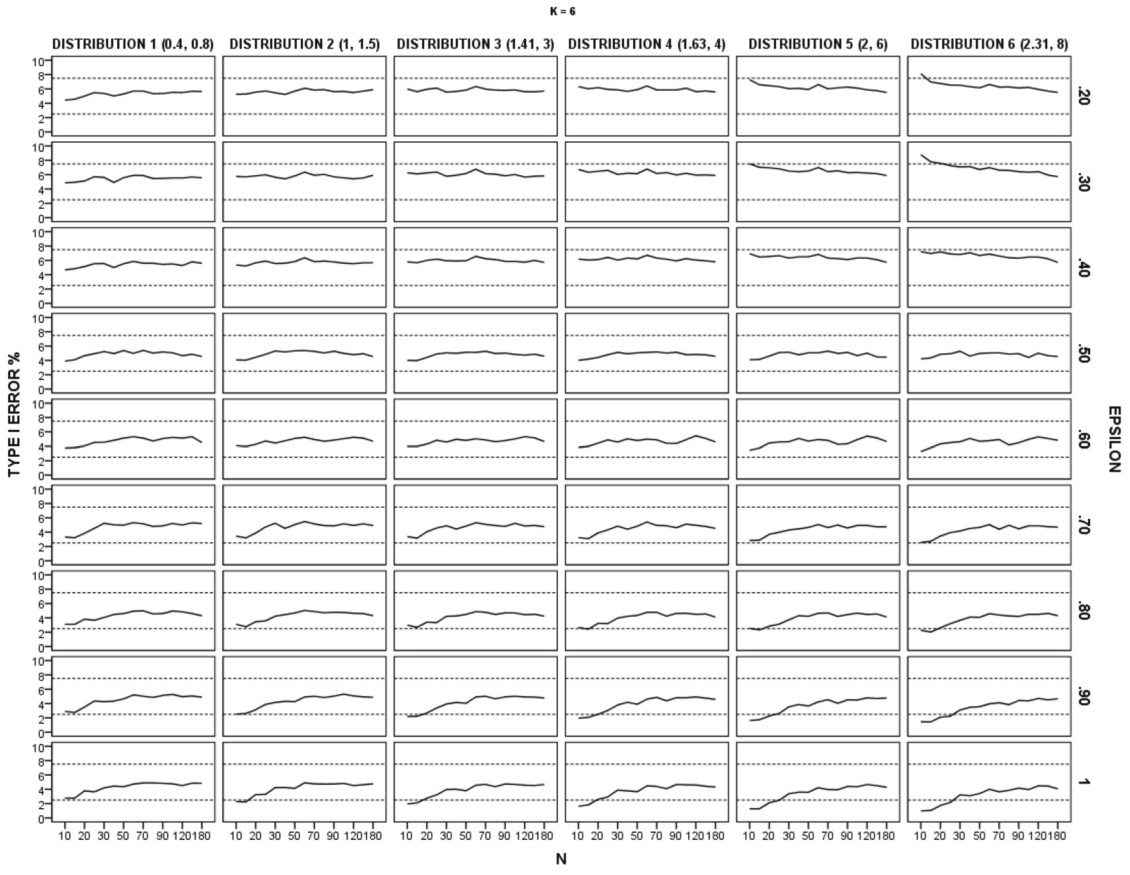
Type I Error Rate (Percentage) for K = 4 as a Function of Distribution Shape, N, and ε̂



Note. In parentheses: skewness and kurtosis coefficients.

Figure 3

Type I Error Rate (Percentage) for K = 6 as a Function of Distribution Shape, N, and $\hat{\varepsilon}$



Note. In parentheses: skewness and kurtosis coefficients.

Table 5

Summary of the Results Obtained for Type I Error

D	γ_1	γ_2	*K* = 3	*K* = 4	*K* = 6
1	0.4	0.8	Robust	Robust	Robust
2	1	1.5	Robust	Robust	C: $\hat{\varepsilon}$= 1, *N* = 10-15 Otherwise robust
3	1.41	3	Robust	Robust	C: $\hat{\varepsilon}$= .90, *N* = 10-15 $\hat{\varepsilon}$= 1, *N* = 10-15 Otherwise robust
4	1.63	4	Robust	Robust	C: $\hat{\varepsilon}$= .80, *N* = 15 $\hat{\varepsilon}$= .90, *N* = 10-15 $\hat{\varepsilon}$= 1, *N* = 10-15 Otherwise robust
5	2	6	C: $\hat{\varepsilon}$= 1, *N* = 10 Otherwise robust	L: $\hat{\varepsilon}$= .33, *N* = 10 $\hat{\varepsilon}$= .40, *N* = 10 C: $\hat{\varepsilon}$= .90, *N* =10 $\hat{\varepsilon}$= 1, *N* = 10 Otherwise robust	L: $\hat{\varepsilon}$= .30, *N* = 10 C: $\hat{\varepsilon}$= .80, *N* = 15 $\hat{\varepsilon}$= .90, *N* = 10-20 $\hat{\varepsilon}$= 1, *N* = 10-25 Otherwise robust
6	2.31	8	L: $\hat{\varepsilon}$= .50, *N* = 10 $\hat{\varepsilon}$= .60, *N* = 10 C: $\hat{\varepsilon}$= 1, *N* = 10-15 Otherwise robust	L: $\hat{\varepsilon}$= .33, *N* = 10 $\hat{\varepsilon}$= .40, *N* = 10 $\hat{\varepsilon}$= .50, *N* = 10 C: $\hat{\varepsilon}$= .90, *N* = 10 $\hat{\varepsilon}$= 1, *N* = 10-25 Otherwise robust	L: $\hat{\varepsilon}$= .20, *N* = 10 $\hat{\varepsilon}$= .30, *N* = 10-20 C: $\hat{\varepsilon}$= .80, *N* = 10-15 $\hat{\varepsilon}$= .90, *N* = 10-25 $\hat{\varepsilon}$= 1, *N* = 10-25 Otherwise robust

Note. D: Distribution; C: Conservative; L: Liberal; γ_1: Skewness; γ_2: Kurtosis.

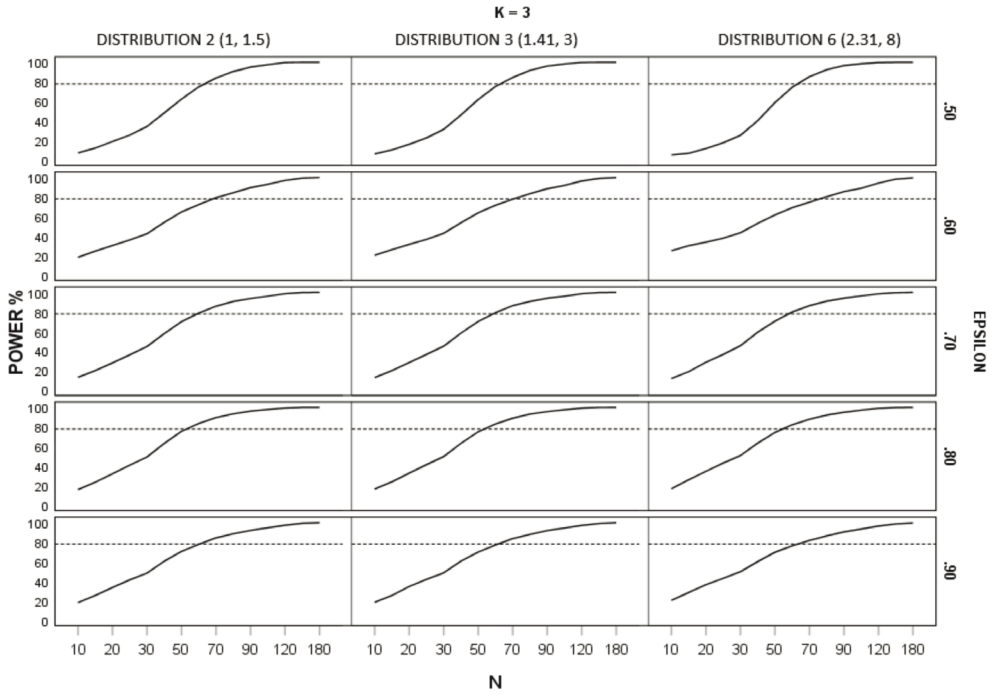
with high values of $\hat{\varepsilon}$ ($\hat{\varepsilon}$ =1 for *K* = 3, and $\hat{\varepsilon} \geq .90$ for *K* = 4) and small sample size (*N* = 10). This tendency to be conservative is also found for both *K* = 3 and *K* = 4 for the distribution with $\gamma_1 = 2.31$ and $\gamma_2 = 8$ for high values of $\hat{\varepsilon}$ and small sample size. However, with this distribution *B-F* tends to be liberal with *N* = 10 and lower values of epsilon ($\hat{\varepsilon} \leq .60$ for *K* = 3 and $\hat{\varepsilon} \leq .50$ for *K* = 4).

For *K* = 6, *B-F* is only robust under all conditions for the distribution with slight deviation from normality ($\gamma_1 = 0.4$, $\gamma_2 = 0.8$). With the remaining distributions, the test tends to be liberal with extreme deviation from normality, lower values of epsilon, and small sample size. For the distribution with $\gamma_1 = 2$ and $\gamma_2 = 6$, *B-F* is liberal with $\hat{\varepsilon}$ = .30 and *N* = 10, whereas in the case of the distribution with $\gamma_1 = 2.31$ and $\gamma_2 = 8$, *B-F* is liberal with $\hat{\varepsilon}$ = .20 and *N* = 10 and with $\hat{\varepsilon}$ = .30 and *N* = 10-20. In addition, and as with *K* = 3 and 4, it tends to be conservative with high values of $\hat{\varepsilon}$ and small sample size. The worst scenario is with extreme deviation from normality ($\gamma_1 = 2.31$, $\gamma_2 = 8$), in which *B-F* is conservative with $\hat{\varepsilon}$ = .80 and *N* = 10-15, and with $\hat{\varepsilon} \geq .90$ and *N* = 10-25.

Statistical Power

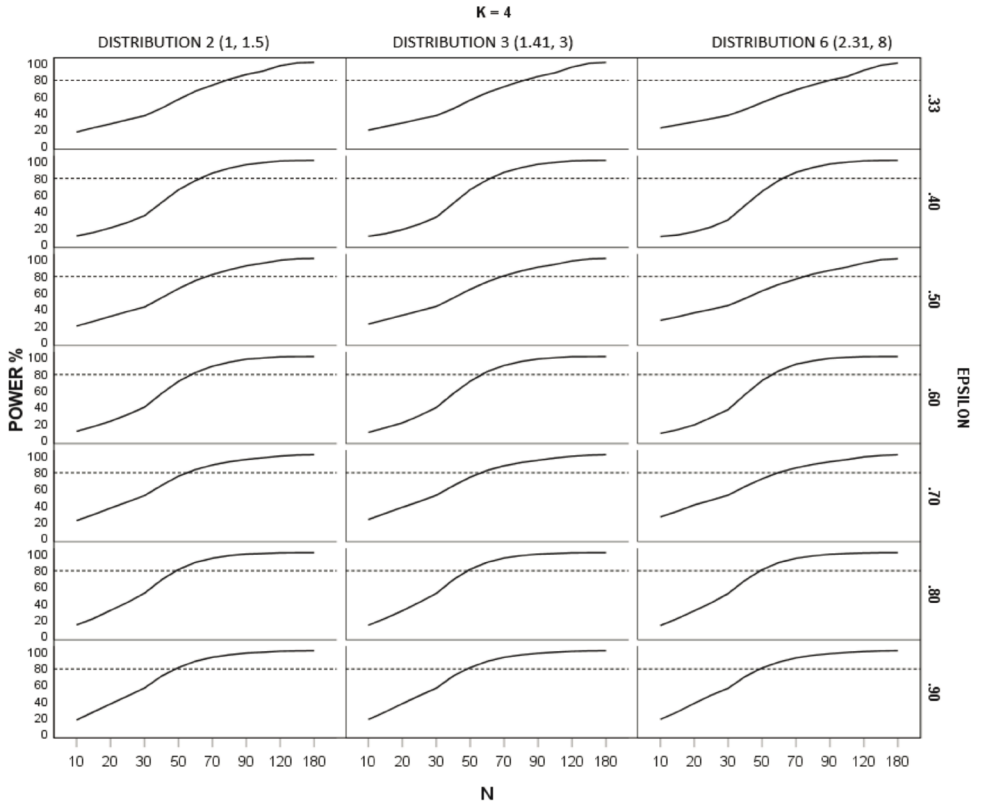
Empirical power for each *K*, distribution, *N*, and $\hat{\varepsilon}$ value are displayed in Figures 4-6 (detailed tables can be found at https://dx.doi.org/10.24310/riuma.37706). Table 6 shows the sample size at which a power of 80% is reached. As expected, the results show

Figure 4
Power (Percentage) for K = 3 as a Function of Distribution Shape, N, and ε̂



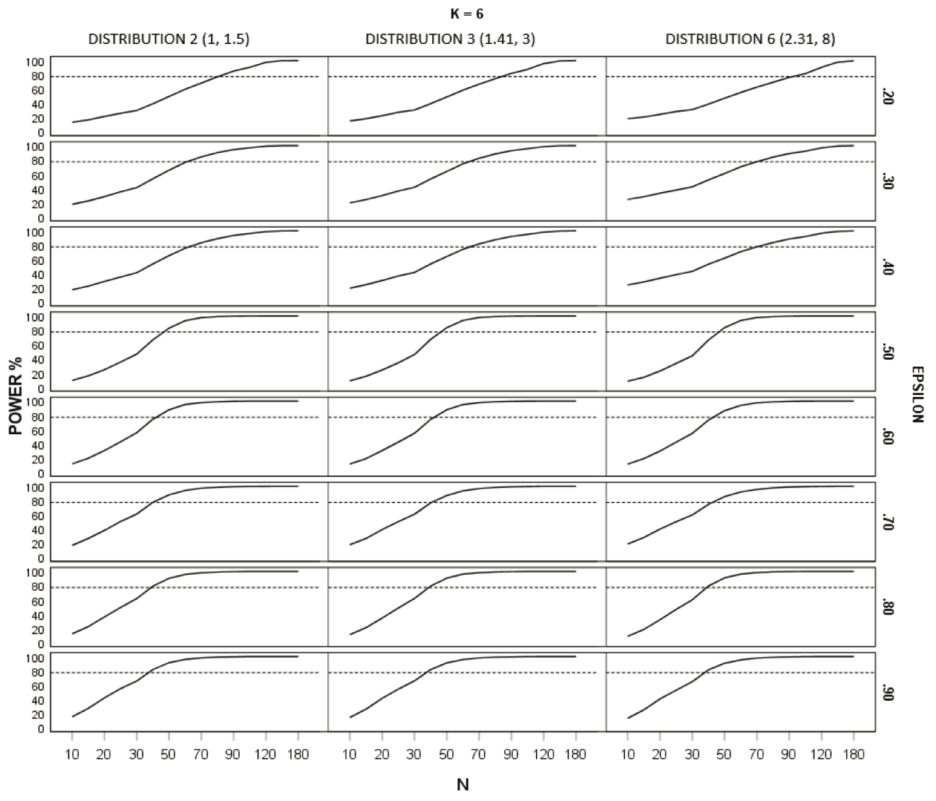
Note. In parentheses: skewness and kurtosis coefficients.

Figure 5
Power (Percentage) for K = 4 as a Function of Distribution Shape, N, and ε̂



Note. In parentheses: skewness and kurtosis coefficients.

Figure 6
Power (Percentage) for K = 6 as a Function of Distribution Shape, N, and $\hat{\varepsilon}$



Note. In parentheses: skewness and kurtosis coefficients.

Table 6
Sample Size at Which a Power of 80% is Reached as a Function of Distribution Shape, $\hat{\varepsilon}$, and Number of Repeated Measures (K)

K	$\hat{\varepsilon}$	Distribution 2 ($\gamma_1 = 1$; $\gamma_2 = 1.5$)	Distribution 3 ($\gamma_1 = 1.41$; $\gamma_2 = 3$)	Distribution 6 ($\gamma_1 = 2.31$; $\gamma_2 = 8$)
3	.50	70	70	70
	.60	80	80	80
	.70	70	70	70
	.80	60	60	60
	.90	70	70	70
4	.33	90	90	100
	.40	70	70	70
	.50	70	80	80
	.60	60	60	60
	.70	60	60	70
	.80	60	60	60
	.90	60	60	60
6	.20	90	90	100
	.30	70	70	80
	.40	70	70	80
	.50	50	50	50
	.60	50	50	50
	.70	50	50	50
	.80	50	50	50
	.90	40	40	40

Note. γ_1: skewness; γ_2: kurtosis.

that power increases with sample size and also that it is affected by $\hat{\varepsilon}$ values, such that a large sample is required to reach adequate power with $\hat{\varepsilon}$ values close to the lower bound. Overall, for $K = 3$, this power is achieved with 60-80 participants across all distributions and $\hat{\varepsilon}$ values. For $K = 4$, this power is reached in a range of 70-100 participants when $\hat{\varepsilon} \leq .50$, and 60-70 when $\hat{\varepsilon} \geq .60$. For $K = 6$, 70-100 participants are required when $\hat{\varepsilon} \leq .40$, and 40-50 for $\hat{\varepsilon} \geq .50$.

Discussion

The aim of this study was to extend knowledge about the robustness and power of *B-F* by considering a wider range of conditions than has been the case previously. To this end, we simulated designs with 3, 4, and 6 repeated measures, sample sizes from 10 to 180, $\hat{\varepsilon}$ values from the corresponding lower bound to 1, and six distributions representing slight to extreme deviations from normality.

Regarding robustness, the results show that Type I error rates vary as a function of the number of repeated measures, distribution shape, epsilon value, and sample size. For $K = 3$, *B-F* is robust for distributions with $\gamma_1 \leq 1.63$ and $\gamma_2 \leq 4$ in all conditions manipulated. However, with higher values of γ_1 and γ_2 the procedure becomes conservative with $\hat{\varepsilon} = 1$ and $N = 10$. With the most extreme deviation from normality ($\gamma_1 = 2.31$ and $\gamma_2 = 8$), the procedure is liberal with smaller values of $\hat{\varepsilon}$ and $N = 10$. These results indicate that for $K = 3$, *B-F* remains robust with violation of both sphericity and normality for distributions with $\gamma_1 \leq 1.63$ and $\gamma_2 \leq 4$, but with more severe deviations from normality a sample size larger than 10 is required for low values of $\hat{\varepsilon}$.

For $K = 4$, *B-F* is again robust for distributions with $\gamma_1 \leq 1.63$ and $\gamma_2 \leq 4$ in all conditions manipulated. With higher values of γ_1 and γ_2 and for $N = 10$ the procedure becomes conservative with $\hat{\varepsilon} = 1$, and liberal with low values of $\hat{\varepsilon}$. Overall, *B-F* is suitable for use with extreme deviation from both normality and sphericity when sample size is larger than 10.

For $K = 6$, *B-F* is robust with very slight deviation from normality ($\gamma_1 = 0.4$ and $\gamma_2 = 0.8$) in all conditions studied. With the other distributions considered, it is conservative with high values of $\hat{\varepsilon}$ and small sample size. A tendency towards liberality appears with severe deviation from normality, $\gamma_1 \geq 2$ and $\gamma_2 \geq 6$, with small values of $\hat{\varepsilon}$ ($\hat{\varepsilon} \leq 30$) and small sample sizes ($N = 10$ and 20). These results indicate that *B-F* may be used under extreme deviation from both normality and sphericity when sample size is larger than 20.

The results regarding liberality of *B-F* appear to contradict those of Berkovits et al. (2000), who found that the procedure was robust under all manipulated conditions. However, their study was conducted under more limited conditions (specifically, $K = 4$ and epsilon > .48) than was the case here. Consistent with Berkovits et al. (2000), our results for $K = 4$ and $\hat{\varepsilon} = .50$ likewise show that *B-F* is robust under all non-normality conditions. Our findings are also in line with those reported by Vallejo et al. (2006, 2010) when using a 3x4 split-plot design and $\hat{\varepsilon} \geq .50$. The tendency we observed for *B-F* to be conservative with high $\hat{\varepsilon}$ values was also documented in both these previous studies.

As a general rule, the first point to consider is that *B-F* may become conservative with higher values of $\hat{\varepsilon}$ (e.g., $\hat{\varepsilon} \geq .80$ for $K = 6$), in which case adjusted *F*-tests, such as Greenhouse-Geisser and Huynh-Feldt adjustments, may be a better option (Blanca et al., 2023b, 2024). Second, *B-F* is suitable for use under violation of both sphericity and normality for distributions with $\gamma_1 \leq 1.63$ and $\gamma_2 \leq 4$. With non-normal distributions of these characteristics, *B-F* is superior to adjusted *F*-tests insofar as the latter have shown a tendency to be liberal with $N = 10$ and low values of $\hat{\varepsilon}$ (Blanca et al., 2024). Third, with more extreme deviation from normality, *B-F* yields reliable results if $N > 20$. More specifically, *B-F* requires $N > 10$ for $K = 3$ if $\hat{\varepsilon} \leq .60$ and for $K = 4$ if $\hat{\varepsilon} \leq .50$, whereas $N > 20$ is required for $K = 6$ if $\hat{\varepsilon} \leq .30$. In these scenarios, *B-F* is slightly superior to adjusted *F*-tests as the latter require $N > 30$ (Blanca et al., 2024).

A possible option in those scenarios where *B-F* is liberal (e.g., under extreme violation of both normality and sphericity and small sample size) is to use a more stringent alpha level. This solution has been proposed previously with other statistical tests (Blanca et al., 2018; Keppel & Wickens, 2004; Tabachnick & Fidell, 2007). Here we conducted simulations of these cases (see Table 5), considering nominal alpha levels of 2.5% and 1%, and computing *B-F* (results are shown in Table 7). In general, a nominal alpha level of 2.5% is sufficient to keep the Type I error rate for *B-F* within [2.5%, 7.5%] in all conditions. It is important to clarify here that using Bradley's liberal criterion implies that the researcher assumes that the actual significance level is between 2.5% and 7.5% for the corresponding nominal value (5%, or 2.5% when a more stringent alpha level is used).

As for empirical power, our results show that power increases with sample size, reflecting the well-known relationship between the two. We also found that deviation from normality did not affect the power of *B-F*. However, it is more sensitive to non-sphericity: the greater the violation of sphericity, with $\hat{\varepsilon}$ values close to the lower bound, the larger the sample size required to ensure adequate power. For example, and assuming 80% power to be adequate (Cooper & Garson, 2016; Kirk, 2013), a sample size of 90-100 is required for $K = 6$ and $\hat{\varepsilon} = .20$, whereas for $\hat{\varepsilon} = .60$, 50 participants are sufficient to reach 80% power for a medium effect size. If we compare these results with those reported by Blanca et al. (2024) for the two adjusted *F*-tests, then *B-F* seems to have greater power in some cases as it is less affected by non-normality.

Table 7
Type I Error Rates for Bootstrap-F (in Percentages) for a Nominal Alpha of 2.5% (1% in Parentheses) in the Conditions Under Which it is not Robust at the 5% Nominal Alpha Level (γ_1: Skewness; γ_2: Kurtosis)

K	$\hat{\varepsilon}$	N	$\gamma_1 = 2, \gamma_2 = 6$	$\gamma_1 = 2.31, \gamma_2 = 8$
3	.50	10		4.62 (2.72)
	.60	10		4.60 (2.48)
4	.33	10	5.00 (3.16)	5.68 (3.82)
	.40	10	5.70 (3.52)	6.42 (4.30)
	.50	10		4.76 (2.82)
6	.20	10		5.46 (3.64)
	.30	10	5.26 (3.40)	6.06 (4.00)
	.30	15		5.26 (3.34)
	.30	20		5.14 (3.06)

In conclusion, the *B-F* procedure offers an alternative for the analysis of repeated measures data with a nominal alpha of 5% under certain conditions specified in Table 5, which researchers may consult to decide if it is a correct option given the characteristics of their data. As a rule of thumb, and to ensure that *B-F* remains robust under non-normality and non-sphericity, a $N > 20$ is required to maintain Type I error rates ≤ 7.5%. In the event of extreme violations of both normality and sphericity and $10 \leq N \leq 20$, *B-F* may be used if a more stringent alpha level (e.g., 2.5%) is considered. It should also be noted that with high $\hat{\varepsilon}$ values the procedure may become conservative and require a $N > 25$. Researchers may consult Table 6 to determine the sample size at which 80% power is reached as a function of the number of repeated measures and other data characteristics.

Researchers may also wish to consider other alternatives to *B-F*, including the adjusted *F*-tests mentioned above, as well as classical non-parametric tests such as the Friedman test, multivariate analysis, and the linear mixed model (LMM). However, simulation studies have shown that these procedures also have limitations and can become liberal with violations of sphericity and small sample sizes (Berkovits et al., 2000; Blanca et al., 2023b, 2024; Harwell & Serlin, 1994; Haverkamp & Beauducel, 2017, 2019; Hayoz, 2007). A further limitation of the LMM relates to problems identifying the true structure of the covariance matrix (Brown & Prescott, 2006). An interesting line of future research would therefore be to compare these procedures and to analyze how they perform when used in conjunction with the bootstrap method.

This study has a number of limitations that need to be acknowledged. First, the results are applicable only to the conditions studied here, that is, to designs containing 3, 4, and 6 repeated measures, and to non-normal distributions with values of skewness and kurtosis coefficients up to 2.31 and 8, respectively. Although

these conditions reflect a wide range of real-life scenarios, future research might focus on exploring the performance of *B-F* in designs with a larger number of repeated measures, in more complex experimental designs that incorporate both within- and between-subject factors, and with distributions showing greater deviation from normality. Investigation of these scenarios will provide a deeper understanding of the applicability of the procedure in various research contexts. Second, we have considered the unstructured covariance matrix as being the most general structure. Further research might include other types of structures that contemplate serial correlation, such as autoregressive, heterogeneous autoregressive, Toeplitz, etc. This would help to extend knowledge about the robustness of *B-F* under different dependency structures. Third, the data simulated here include complete cases without accounting for the presence of missing values. The importance of detecting patterns of missing data and mechanisms of loss, as well as selecting an appropriate imputation method, is widely acknowledged (Berglund & Heeringa, 2014; Vallejo et al., 2011). A possible avenue for further research would therefore be to analyze both Type I error and power of *B-F* with different patterns of missing data and different imputation methods. Finally, the present study focuses on the comparison of untrimmed means, so it would be interesting to explore the performance of *B-F* with trimmed means. Outliers often pose difficulties in data analysis, and the use of trimmed means is a procedure that can deal with this problem (Wilcox, 2022).

Author Contributions

María J. Blanca: Conceptualization, Methodology, Writing – Original draft, Formal Analysis. **Roser Bono**: Methodology, Software, Writing – Review and Editing. **Jaume Arnau**: Software, Writing – Review and Editing. **F. Javier García-Castro**: Methodology, Writing – Review and Editing. **Rafael Alarcón**: Methodology, Formal Analysis, Writing – Review and Editing. **Guillermo Vallejo**: Software, Writing – Review and Editing.

Acknowledgements

The authors would like to thank Macarena Torrado for her collaboration in this study.

Funding

This research was supported by the Ministry of Science and Innovation (grant PID2020-113191GB-I00 from the MCIN/ AEI/ 10.13039/501100011033 and by funding from the Regional Government of Andalusia to Consolidated Research Group CTS110). This funding source had no role in the design of this study, data collection, management, analysis, and interpretation of data, writing of the manuscript, or the decision to submit the manuscript for publication.

Declaration of Interests

The authors declare that there are no conflicts of interest.

Data Availability Statement

Data are available at https://dx.doi.org/10.24310/riuma.37706

References

Arnau, J., Bendayan, R., Blanca, M. J., & Bono, R. (2014). Should we rely on the Kenward–Roger approximation when using linear mixed models if the groups have different distributions? *British Journal of Mathematical and Statistical Psychology*, *67*(3), 408–429. https://doi.org/10.1111/bmsp.12026

Arnau, J., Bono, R., Blanca, M. J., & Bendayan, R. (2012). Using the linear mixed model to analyze nonnormal data distributions in longitudinal designs. *Behavior Research Methods, 44*(4), 1224–1238. https://doi.org/10.3758/s13428-012-0196-y

Berglund, P., & Heeringa, S. (2014). *Multiple imputation of missing data using SAS*. SAS Institute Inc.

Berkovits, I., Hancock, G., & 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. https://doi.org/10.1177/00131640021970961

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? *Behavior Research Methods, 50*(3), 937-962. https://doi.org/10.3758/s13428-017-0918-2

Blanca, M. J., Arnau, J., García-Castro, F. J., Alarcón, R., & Bono, R. (2023a). Non-normal data in repeated measures: Impact on Type I error and power. *Psicothema*, *35*(1), 21–29. https://doi.org/10.7334/psicothema2022.292

Blanca, M. J., Arnau, J., García-Castro, F. J., Alarcón, R., & Bono, R. (2023b). Repeated measures ANOVA and adjusted *F*-tests when sphericity is violated: Which procedure is best? *Frontiers in Psychology, 14*, Article 1192453. https://doi.org/10.3389/fpsyg.2023.1192453

Blanca, M. J., Alarcón, R., Arnau, J., García-Castro, F. J., & Bono, R. (2024). How to proceed when both normality and sphericity are violated in repeated measures ANOVA. *Anales de Psicología / Annals of Psychology*, *40*(3), 466–480. https://doi.org/10.6018/analesps.594291

Bono, R., Arnau, J., & Vallejo, G. (2010). Modelización de diseños split-plot y estructuras de covarianza no estacionarias: un estudio de simulación [Modeling split-plot data and nonstationary covariance structures: A simulation study]. *Escritos de Psicología / Psychological Writings, 3*(3), 1–7. https://doi.org/10.5231/Psy.Writ.2010.2903

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems II. Effect of inequality of variance and of correlation of error in the two-way classification. *Annals of Mathematical Statistics, 25*(3), 484–498. https://doi.org/10.1214/aoms/1177728717

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Brown, H., & Prescott, R. (2006). *Applied mixed models in medicine* (2nd edition). Wiley.

Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researchers* (2nd ed.). John Wiley & Sons, Inc.

Chernick, M. R., & LaBudde, R. A. (2011). *An introduction to bootstrap methods with applications to R*. John Wiley & Sons, Inc.

Christensen, A. P., & Golino, H. (2021). Estimating the stability of psychological dimensions via bootstrap exploratory graph analysis: A Monte Carlo simulation and tutorial. *Psych*, *3*(3), 479–500. https://doi.org/10.3390/psych3030032

Cooper, J. A., & Garson, G. D. (2016). *Power analysis*. Statistical Associates Blue Book Series.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*(1), 1–26. http://www.jstor.org/stable/2958830

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician, 37*(1), 36-48. https://doi.org/10.2307/2685844

Efron, B., & Tibshirani, R. J., (1993). *An introduction to the bootstrap*. Chapman & Hall.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*(4), 521–532. https://doi.org/10.1007/BF02293811

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's results on the use of the *F* distribution in multivariate analysis. *The Annals of Mathematical Statistics*, *29*(3) 885–891. https://doi.org/10.1214/aoms/1177706545

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika 24*(2), 95–112. https://doi.org/10.1007/BF02289823

Haverkamp, N., & Beauducel, A. (2017). Violation of the sphericity assumption and its effect on Type-I error rates in repeated measures ANOVA and multi-level linear models (MLM). *Frontiers in Psychology*, *8*, Article 1841. https://doi.org/10.3389/fpsyg.2017.01841

Haverkamp, N., & Beauducel, A. (2019). Differences of Type I error rates for ANOVA and multilevel-linear-models using SAS and SPSS for repeated measures designs. *Meta-Psychology*, *3,* Article MP.2018.898. https://doi.org/10.15626/mp.2018.898

Harwell, M. R., & Serlin, R. C. (1994). A Monte Carlo study of the Friedman test and some competitors in the single factor, repeated measures design with unequal covariances. *Computational Statistics & Data Analysis*, *17*(1), 35–49. https://doi.org/10.1016/0167-9473(92)00060-5

Hayoz, S. (2007). Behavior of nonparametric tests in longitudinal design. *15th European young statisticians meeting*. http://matematicas.unex.es/~idelpuerto/WEB_EYSM/Articles/ch_stefanie_hayoz_art.pdf

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Publications.

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*(1), 69–82. https://doi.org/10.2307/1164736

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Prentice Hall.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1999). A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology, 52*(1), 63–78. https://doi.org/10.1348/000711099158964

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*(2), 275–298. https://doi.org/10.1111/j.2044-8317.1996.tb01089.x

Kherad-Pajouh, S., & Renaud, O. (2015). A general permutation approach for analyzing repeated measures ANOVA and mixed-model designs. *Statistical Papers*, *56*(4), 947–967. https://doi.org/10.1007/s00362-014-0617-3

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Sage Publications.

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. https://doi.org/10.1177/0013164403260196

Livacic-Rojas, P., Vallejo, G., & Fernández, P. (2010). Analysis of Type I error rates of univariate and multivariate procedures in repeated measures designs. *Communications in Statistics – Simulation and Computation, 39*(3), 624–664. https://doi.org/10.1080/03610910903548952

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods, 52*, 464–488. https://doi.org/10.3758/s13428-019-01246-w

Muller, K., Edwards, L., Simpson, S., & Taylor, D. (2007). Statistical tests with accurate size and power for balanced linear mixed models. *Statistics in Medicine*, *26*(19), 3639–3660. https://doi.org/10.1002/sim.2827

Oberfeld, D., & Franke, T. (2013). Evaluating the robustness of repeated measures analyses: The case of small sample sizes and nonnormal data. *Behavior Research Methods*, *45*(3), 792–812. https://doi.org/10.3758/s13428-012-0281-2

Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental design using ANOVA*. Thomson Brooks/Cole.

Vallejo, G., Ato, M., Fernández, P., & Livacic-Rojas, P. (2013). Multilevel bootstrap analysis with assumptions violated. *Psicothema, 25*(4), 520-528. https://doi.org/10.7334/psicothema2013.58

Vallejo, G., Cuesta, M., Fernández, M., & Herrero, F. (2006). A comparison of the bootstrap-*F*, improved general approximation, and Brown-Forsythe multivariate approaches in a mixed repeated measures design. *Educational and Psychological Measurement, 66*(1), 35–62. https://doi.org/10.1177/0013164404273943

Vallejo, G., Fernández, M. P., Livacic-Rojas, P. E., & Tuero-Herrero, E. (2011) Comparison of modern methods for analyzing repeated measures data with missing values. *Multivariate Behavioral Research, 46*(6), 900–937. https://doi.org/10.1080/00273171.2011.625320

Vallejo, G., Fernández, M. P., Tuero, E., & Livacic-Rojas, P. E. (2010). Análisis de medidas repetidas usando métodos de remuestreo [Analyzing repeated measures using resampling methods]. *Anales de Psicología / Annals of Psychology*, *26*(2), 400–409.

Voelkle, M. C., & McKnight, P. E. (2012). One size fits all? A Monte-Carlo simulation on the relationship between repeated measures (M) ANOVA and latent curve modeling. *Methodology, 8*(1), 23–38. https://doi.org/10.1027/1614-2241/a000044

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. Gulf Professional Publishing.

Wilcox, R. R. (2022). *Introduction to robust estimation and hypothesis testing*. Academic Press.