anales de psicología / annals of psychology
2024, vol. 40, nº 3 (october), 466-480
https://doi.org/10.6018/analesps.594291

ISSN online: 1695-2294.
https://revistas.um.es/analesps

How to proceed when both normality and sphericity are violated in repeated measures ANOVA

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

[1]*Department of Psychobiology and Behavioral Sciences Methodology, University of Malaga*
[2]*Department of Social Psychology and Quantitative Psychology, University of Barcelona,*
[3]*Department of Psychology, Universidad Loyola Andalucía*
[4]*Institute of Neurosciences, University of Barcelona*

Título: Cómo proceder cuando se violan la normalidad y la esfericidad en el ANOVA de medidas repetidas.

Resumen: Las pruebas *F* ajustadas se han propuesto como alternativa al estadístico *F* en el ANOVA de medidas repetidas. A pesar de existir investigación previa, falta evidencia sobre el comportamiento de estos estadísticos en caso de violación simultánea de normalidad y esfericidad. El objetivo del presente trabajo ha sido realizar un examen detallado del error de tipo I y la potencia del estadístico *F* y los ajustes de Greenhouse-Geisser (*F-GG*) y Huynh-Feldt (*F-HF*), manipulando el número de medidas repetidas (3-6), el tamaño de la muestra (10-300), la esfericidad (estimador Greenhouse-Geisser de épsilon, $\hat{\varepsilon}$, desde su límite inferior al superior), y la forma de la distribución (desde desviaciones leves a extremas de la normalidad). Los resultados muestran que el comportamiento de *F-GG* y *F-HF* depende del grado de violación de la normalidad, esfericidad y tamaño muestral. En general, se sugiere utilizar *F-GG* en caso de violación de la esfericidad y desviaciones leves o moderadas de la normalidad; con desviaciones graves de ambos, *F-GG* puede utilizarse con un tamaño muestral superior a 10; y con desviaciones extremas, este estadístico puede utilizarse con un tamaño muestral superior a 30. En caso de resultados discrepantes entre *F-GG* y *F-HF*, la elección depende del valor $\hat{\varepsilon}$.

Palabras clave: Ajuste Greenhouse-Geisser. Ajuste Huynh-Feldt. Robustez. Potencia. Simulación Monte Carlo.

Abstract: Adjusted *F*-tests have typically been proposed as an alternative to the *F*-statistic in repeated measures ANOVA. Despite considerable research, it remains unclear how these statistics perform under simultaneous violation of normality and sphericity. Accordingly, our aim here was to conduct a detailed examination of Type I error and power of the *F*-statistic and the Greenhouse-Geisser (*F-GG*) and Huynh-Feldt (*F-HF*) adjustments, manipulating the number of repeated measures (3-6), sample size (10-300), sphericity (Greenhouse-Geisser epsilon estimator, $\hat{\varepsilon}$, from its lower to upper limit), and distribution shape (slight to extreme deviations from normality). The findings show that the behavior of *F-GG* and *F-HF* depends on the degree of violation of both normality, sphericity, and sample size. Overall, we suggest using *F-GG* under violation of sphericity and slight or moderate deviations from normality in all sample size; with severe deviations from both normality and sphericity *F-GG* may be used with a sample size larger than 10; and with extreme deviation from both normality and sphericity this statistic may be used with a sample size larger than 30. In the event of discrepant results between *F-GG* and *F-HF*, the choice depends on the $\hat{\varepsilon}$ value.

Keywords: Greenhouse-Geisser adjustment. Huynh-Feldt adjustment. Robustness. Power. Monte Carlo simulation.

Introduction

The one-way repeated measures or within-subject design represents situations in which the dependent variable is repeatedly observed under different experimental conditions or at various time points. In this scenario, the conventional statistical procedure based on the general linear model is analysis of variance (RM-ANOVA), which uses the *F*-statistic to test the statistical significance associated with the null hypothesis of equality of means. For a valid statistical decision, this test requires fulfillment of the assumptions of normality and sphericity. Under violations of these assumptions, a number of alternatives have been proposed, including non-parametric procedures, multivariate analysis, use of the linear mixed model, robust statistics or bootstrap methods (Arnau et al., 2012, 2013; Livacic-Rojas et al., 2010; Sheskin, 2003; Wilcox, 2022). However, research has shown that in several areas of knowledge, RM-ANOVA is much more widely used than are these alternatives (e.g., Armstrong, 2017; Blanca et al., 2018; Goedert et al., 2013). In other words, although more sophisticated statistical analyses exist, most applied researchers continue to use RM-ANOVA, probably because it is widely regarded as being easy to apply and simple to interpret.

Monte Carlo simulation studies are useful for analyzing the degree to which the violation of its underlying assumptions affect the Type I error and power of the *F*-test. Regarding normality, the meta-analysis by Keselman et al. (1996) found that the *F*-statistic is generally insensitive to violations of normality, a result that is in line with other research (e.g., Berkovits et al., 2000; Kherad-Pajouh & Renaud, 2015). More recently, Blanca et al. (2023a) carried out an exhaustive simulation study, manipulating the number of repeated measures (3, 4, 6, and 8), sample size (from 10 to 300), and distribution shape (slight, moderate, and severe departure from normality). Their results showed, consistent with the previous evidence, that Type I error and power are not affected by violations of normality as long as sphericity is met.

The violation of sphericity is known to have a more severe impact than non-normality on robustness of the *F*-statistic, inflating Type I error (e.g., Berkovits et al., 2000; Haverkamp & Beauducel, 2017, 2019; Voelkle & McKnight, 2012). One of the procedures for controlling Type I error involves reducing the degrees of freedom of the *F*-statistic by a multiplicative factor called epsilon (ε), as a result of which it becomes a more demanding test (Box, 1954). The

*** Correspondence address [Dirección para correspondencia]:**
María J. Blanca. Facultad de Psicología y Logopedia. C/. Doctor Ortiz Ramos, 12. Ampliación de Teatinos. 29010-Málaga (Spain).
E-mail: blamen@uma.es
(Article received: 23-11-2023; revised: 08-01-2024; accepted: 24-01-2024)

value of ε represents the amount by which the data depart from sphericity, and it ranges from $1/K$-1 to 1, where K is the number of repeated measurements. Sphericity is satisfied if ε is equal to 1. The further ε departs from 1 and the closer it approaches its lower limit the greater the violation of the assumption. Tests using reduced degrees of freedom are known as adjusted F-tests, two of which are widely used and available in most statistical software: the Greenhouse-Geisser adjusted F-test (*F-GG*; Box, 1954; Geisser & Greenhouse, 1958; Greenhouse & Geisser, 1959), whose ε estimator is known as $\hat{\varepsilon}$, and the Huynh-Feldt adjusted F-test (*F-HF*; Huynh & Feldt, 1976), whose ε estimator is referred to as $\tilde{\varepsilon}$.

Simulation studies exploring sphericity violation with normal data and a one-way design have yielded inconsistent results. Some have found that both *F-HF* and *F-GG* are robust to sphericity violations (Berkovits et al., 2000; Muller et al., 2007), whereas others report that *F-HF* outperforms *F-GG*, especially with a large number of repeated measures and small sample size (Haverkamp & Beauducel, 2017, 2019; Oberfeld & Franke, 2013). These results contrast with other research and with what is stated in some classic methodological books, in which the use of *F-GG* is recommended over *F-HF* (Kirk, 2013; Maxwell & Delaney, 2004; Voelkle & McKnight, 2012). In view of these different recommendations, Blanca et al. (2023b), taking Greenhouse-Geisser $\hat{\varepsilon}$ as a reference, compared the performance of the F-statistic, *F-GG*, and *F-HF* in terms of Type I error and power for different values of $\hat{\varepsilon}$ (ranging from the lower to its upper limit), with 3, 4, and 6 repeated measures and sample size between 10 and 300. For the interpretation of robustness, they used Bradley's (1978) criteria, both liberal and stringent. According to the former criterion, a test is robust if Type I error is between 2.5 and 7.5, while under the latter it is robust if Type I error is between 4.5 and 5.5, in both cases for a significance level of 5%. The results showed that the F-statistic was liberal with values of $\hat{\varepsilon}$ below .70. With $\hat{\varepsilon}$ of .70 and .80, the Type I error remained within Bradley's liberal limits, but was slightly inflated (6-7%) compared with the two adjusted F-tests. With $\hat{\varepsilon}$ of .90, the Type I error was around 5%. By contrast, *F-GG* and *F-HF* were robust across all sphericity violation conditions, although *F-HF* showed slightly greater empirical power, in line with previous research (Algina & Keselman, 1997). The use of the stringent criterion helped Blanca et al. (2023b) to establish a rule-of-thumb in the event of discrepant results from the two procedures. Specifically, they recommend using *F-GG* for $\hat{\varepsilon}$ values below .60, and *F-HF* for $\hat{\varepsilon}$ values equal to or above .60.

Other studies have focused on the performance of several procedures when both normality and sphericity assumptions are simultaneously violated. For instance, Berkovits et al. (2000) simulated data from a one-way design with four repeated measures, with small sample sizes (N = 10, 15, 30, and 60), non-normal distributions with different values of skewness (γ_1) and kurtosis (γ_2) (1, .75; 1.75, 3.75; and 3, 21, respectively), and different values of ε (.48, .57, .75, and 1). They found that as skewness and kurtosis increased, and with sample sizes equal to or less than 30, *F-GG* and *F-HF* could be conservative with ε of 1 and .75, but liberal with ε of .57 and .48. At sample sizes of 60, *F-GG* and *F-HF* were robust to all violations of normality and sphericity. Oberfeld and Franke (2013) included designs with 4, 8, and 16 repeated measures, lognormal and chi-square distributions with two degrees of freedom, different structures of the covariance matrices with ε equal to .50 and 1, and sample sizes between 3 and 100. With non-normal data, no pattern was found that defined the performance of *F-GG* and *F-HF*. Both could be conservative or liberal, depending on the sample size, number of repeated measures, and type of covariance matrix.

In summary, the results from simulation studies suggest that: a) non-normality does not affect the F-statistic as long as sphericity is met; b) sphericity, irrespective of normality, has serious consequences on the test's robustness, although the two adjusted F-tests may be valid alternatives; and c) there are no clear guidelines when simultaneous violations of normality and sphericity occur, as the impact seems to depend on other factors, such as the degree of sphericity violation and sample size. The purpose of the present study was therefore to conduct a detailed examination of the Type I error and power of the F-statistic, *F-GG*, and *F-HF* under a greater number of conditions than have been analyzed in previous studies, including different numbers of repeated measures, a wide range of non-normal distributions and sphericity violations, and small, medium and large sample sizes. For Type I error, we analyze 4807 conditions, with K = 3, 4, and 6, and including sample sizes from 10 to 300, values of $\hat{\varepsilon}$ from its lower limit to .90 as a function of K, and 11 distributions from slight to extreme deviations from normality. For power analysis, we analyze 3040 conditions, considering designs with K = 3, 4, and 6, two mean patterns for each K, and four non-normal distributions (slight, moderate, severe, and extreme). Our ultimate goal with this study was to clarify the conditions under which the above statistics can be used in the event of simultaneous violation of normality and sphericity.

Methods

A simulation study was carried out using the interactive matrix language (IML) module of SAS 9.4. Data were generated using a series of macros constructed for this purpose. For the generation of non-normal data we used the procedure proposed by Fleishman (1978), which applies a polynomial transformation that simulates data with specific values of skewness and kurtosis. Unstructured covariance matrices with different values of $\hat{\varepsilon}$ were generated. The probability of the values associated with the F-statistic, *F-GG,* and *F-HF* was obtained using PROC GLM of SAS. For each condition, we performed ten thousand replications.

Type I Error

The variables manipulated for a one-way design were as follows:

1. Number of repeated measures (*K*): The repeated measures were *K* = 3, 4, and 6.
2. Total sample size: The sample sizes considered were 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100, 120, 150, 180, 210, 240, 270, and 300, a range that covers small, medium, and large samples.
3. Epsilon ($\hat{\varepsilon}$): Values ranged approximately, depending on the number of repeated measures, between the lower limit and values close to 1 (*K* = 3: .50, .60, .70, .80, .90; *K* = 4: .33, .40, .50, .60, .70, .80, .90; and *K* = 6: .20, .30, .40, .50, .60, .70, .80, .90). These $\hat{\varepsilon}$ values were estimated following the Greenhouse-Geisser procedure (Box, 1954; Geisser & Greenhouse, 1958; Greenhouse & Geisser, 1959).
4. Shape of the distribution: Eleven different distributions were used, both known and unknown, chosen from among those considered by Blanca et al. (2023a), with skewness and kurtosis values ranging from slight to extreme deviations from normality. Table 1 shows their characteristics. Blanca et al. (2013) found that 80% of real data presented values of skewness and kurtosis ranging between -1.25 and 1.25. Distributions 1-5 were selected based on this finding. Distributions 6-11 were included to represent well-known distributions with more severe departure from normal distribution which have been typically used in simulation studies and are also representative of real data (Bono et al., 2017; Micceri, 1989).

Table 1
Skewness (γ_1) and kurtosis (γ_2) coefficients for each simulated distribution.

Distribution	Type	γ_1	γ_2
1	-	0	0.8
2	-	0.8	0
3	-	0.4	0.8
4	-	0.8	0.4
5	-	1	1
6	Gamma (α = 4)	1	1.50
7	Double exponential	0	3
8	Chi-squared (8 d.f.)	1	3
9	Gamma (α = 2)	1.41	3
10	Exponential	2	6
11	Gamma (α = 0.75)	2.31	8

We recorded Type I error rates, which reflect the percentage rejection of the null hypothesis when the differences between the means of the repeated measures are set to zero at the 5% significance level. Bradley's (1978) liberal criterion was used to interpret the results, according to which a procedure is robust if the Type I error rate is between 2.5% and 7.5% for a nominal alpha of 5%. We also considered Bradley's (1978) stringent criterion, whereby a procedure is robust if the Type I error rate is between 4.5% and 5.5% for a nominal alpha of 5%. If the Type I error rate is below the respective lower limit, the procedure is considered conservative, while if it is above the respective upper limit, it is considered liberal.

Empirical Power

To analyze empirical power, we selected mean values to give a medium effect size, $f \approx 0.25$. The number of repeated measures (*K*), sample sizes, and epsilon values ($\hat{\varepsilon}$) were the same as those for Type I error. The other variables manipulated were as follows:

1. Pattern of means: For all repeated measures (*K* = 3, 4, and 6) we used a linear pattern in which the means increase linearly and proportionally to each other (e.g., 1, 1.25, 1.50, 1.75). In addition, for *K* = 3 we used a pattern of means in which one of the means was different from the means of the other repeated measures (e.g., 0, 0, 1). With *K* = 4 and 6, we also used a pattern in which half of the means were different and equal to each other (e.g., 0, 0, 1, 1; 0, 0, 0, 1, 1, 1).
2. Shape of the distribution: Distributions 3, 6, 9 and 11 (see Table 1) were chosen, representing slight (γ_1 = .4 and γ_2 =.8), moderate (γ_1 = 1, γ_2 = 1.50), severe (γ_1 = 1.41, γ_2 = 3), and extreme deviation from normality (γ_1 = 2.31, γ_2 = 8).

Empirical power was calculated using the non-centrality parameter for each pattern of means at a significance level of 5%. The non-centrality parameter is the distance between the distributions of the null and the alternative hypothesis. For power calculations, we used the expected values of the epsilon estimator for the Greenhouse-Geisser and Huynh-Feldt tests to compute the degrees of freedom for the non-central *F* (Muller & Barton, 1989).

Results

Empirical Type I Error Rates

In order to summarize the results, descriptive statistics for empirical Type I error rates were collapsed for all *K* (3, 4, and 6) with distribution shapes and epsilon values that showed the same behavior for the *F*-statistic, *F-GG,* and *F-HF*. Tables 2-3 display the results found with distributions 1-8, Tables 4-5 with distribution 9, and Tables 6-7 with distributions 10-11. Table 8 shows the results for epsilon equal to .90 for all distributions. Table 9 displays the percentage robustness of the *F*-statistic, *F-GG,* and *F-HF* according to Bradley's (1978) stringent criterion. Detailed tables are available as supplementary material.

Table 2

Type I error rates (in percentages) for the F-statistic, F-GG, and F-HF by N across distributions 1-8 ($\gamma_1 \leq 1$, $\gamma_2 \leq 3$) for all K (3, 4, and 6) and $\hat{\varepsilon} \leq .60$. Type I error rates > 7.5 are in bold (liberal).

N	*F*				*F-GG*				*F-HF*			
	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*
10	7.24	**16.28**	**10.55**	2.34	3.20	6.74	5.18	0.93	4.18	7.44	6.03	0.78
15	7.20	**15.98**	**10.18**	2.13	3.90	6.68	5.20	0.72	4.36	6.94	5.70	0.60
20	7.20	**15.50**	**10.05**	2.12	3.52	6.40	5.23	0.65	4.42	6.80	5.58	0.56
25	6.80	**15.74**	**9.91**	2.12	3.72	6.30	5.09	0.57	4.14	6.50	5.37	0.51
30	7.10	**14.63**	**9.91**	1.95	4.18	6.18	5.20	0.47	4.26	6.36	5.42	0.42
40	7.28	**14.64**	**9.73**	1.96	4.20	6.02	5.07	0.41	4.20	6.22	5.23	0.38
50	6.82	**14.75**	**9.68**	1.96	4.18	6.06	5.07	0.41	4.22	6.12	5.19	0.39
60	7.04	**14.64**	**9.62**	1.94	4.08	5.76	5.02	0.38	4.34	6.04	5.12	0.36
70	7.24	**14.58**	**9.67**	1.89	4.34	6.08	5.06	0.33	4.48	6.08	5.15	0.32
80	6.72	**14.22**	**9.57**	1.85	4.18	5.88	5.05	0.36	4.34	5.96	5.12	0.34
90	6.84	**14.28**	**9.61**	1.78	4.52	5.68	5.05	0.27	4.56	5.82	5.12	0.27
100	6.58	**14.34**	**9.58**	1.90	4.10	5.88	5.05	0.35	4.14	5.96	5.11	0.34
120	6.96	**14.60**	**9.52**	1.86	4.14	5.66	4.96	0.34	4.22	5.76	5.02	0.33
150	6.54	**14.26**	**9.50**	1.86	4.04	6.02	4.99	0.38	4.08	6.11	5.03	0.37
180	6.96	**14.02**	**9.53**	1.79	4.08	5.80	4.99	0.31	4.08	5.80	5.03	0.31
210	6.76	**13.88**	**9.45**	1.82	3.96	5.72	4.94	0.34	3.96	5.72	4.97	0.34
240	6.92	**13.99**	**9.53**	1.75	4.20	5.74	5.00	0.32	4.22	5.76	5.02	0.32
270	6.72	**13.98**	**9.47**	1.84	4.04	5.80	4.98	0.32	4.10	5.82	5.01	0.31
300	6.67	**14.46**	**9.49**	1.94	4.34	6.02	4.97	0.31	4.42	6.02	4.99	0.31
Total	6.54	**16.28**	**9.71**	1.95	3.20	6.74	5.06	0.47	3.96	7.44	5.22	0.50

Table 3

Type I error rates (in percentages) for the F-statistic, F-GG, and F-HF by N across distributions 1-8 ($\gamma_1 \leq 1$, $\gamma_2 \leq 3$) for all K (3, 4, and 6) and $\hat{\varepsilon} = .70$ and .80. Type I error rates > 7.5 are in bold (liberal).

N	*F*				*F-GG*				*F-HF*			
	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*
10	5.46	**7.68**	6.56	0.60	2.50	5.08	3.83	0.72	4.36	6.12	5.18	0.41
15	5.48	**7.78**	6.58	0.54	3.32	5.24	4.26	0.54	4.48	5.90	5.16	0.33
20	5.46	**8.14**	6.55	0.51	3.40	5.34	4.42	0.51	4.38	5.76	5.13	0.35
25	5.54	**7.36**	6.43	0.49	3.60	5.44	4.46	0.41	4.40	5.62	4.99	0.30
30	5.48	**7.86**	6.52	0.51	3.66	5.28	4.60	0.45	4.26	5.68	5.06	0.34
40	5.36	**7.84**	6.55	0.59	3.78	5.50	4.70	0.37	4.28	5.76	5.04	0.36
50	5.52	**7.82**	6.54	0.58	4.00	5.54	4.72	0.36	4.32	5.82	5.03	0.36
60	5.36	**7.74**	6.47	0.56	4.18	5.40	4.75	0.27	4.50	5.62	4.96	0.27
70	5.32	**7.64**	6.51	0.55	3.94	5.62	4.79	0.38	4.14	5.76	5.00	0.32
80	5.38	**7.56**	6.57	0.52	4.04	5.50	4.86	0.31	4.26	5.64	5.04	0.32
90	5.68	7.30	6.45	0.45	3.84	5.48	4.79	0.35	4.08	5.62	4.95	0.34
100	5.40	**7.58**	6.43	0.50	4.24	5.36	4.79	0.28	4.34	5.43	4.92	0.27
120	5.54	**7.70**	6.49	0.46	4.32	5.50	4.86	0.30	4.40	5.66	4.98	0.29
150	4.94	**7.56**	6.38	0.59	3.92	5.42	4.76	0.32	3.94	5.48	4.84	0.34
180	5.24	**7.78**	6.51	0.54	4.28	5.58	4.86	0.29	4.36	5.58	4.95	0.30
210	5.58	7.32	6.45	0.44	4.28	5.42	4.84	0.26	4.36	5.42	4.91	0.26
240	5.66	**7.60**	6.51	0.46	4.38	5.54	4.93	0.29	4.44	5.58	4.98	0.29
270	5.77	**7.90**	6.49	0.52	4.18	5.86	4.92	0.36	4.30	5.96	4.97	0.37
300	5.38	7.34	6.45	0.54	4.36	5.56	4.90	0.28	4.42	5.68	4.95	0.29
Total	4.94	**8.14**	6.50	0.52	2.50	**7.60**	4.69	0.47	3.94	**7.62**	5.00	0.34

Table 4

Type I error rates (in percentages) for the F-statistic, F-GG, and F-HF by N with distribution 9 ($\gamma_1 = 1.43, \gamma_2 = 3$) for all K (3, 4, and 6) and $\hat{\varepsilon} \leq .60$. Type I error rates > 7.5 are in bold (liberal).

N	F				F-GG				F-HF			
	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*
10	**8.67**	**17.31**	**11.64**	2.67	3.99	**7.88**	6.32	1.35	5.96	**8.73**	7.26	0.88
15	**8.45**	**16.26**	**11.02**	2.51	3.88	7.14	6.03	1.08	4.93	7.42	6.63	0.75
20	**7.89**	**15.96**	**10.65**	2.44	3.98	7.03	5.92	0.94	4.91	7.08	6.33	0.67
25	**7.77**	**14.99**	**10.47**	2.09	4.78	6.66	5.92	0.61	5.48	6.81	6.21	0.47
30	**7.84**	**15.26**	**10.45**	2.32	4.50	6.68	5.75	0.70	5.04	6.68	6.00	0.55
40	**7.52**	**14.93**	**10.06**	2.19	4.32	6.41	5.50	0.57	4.72	6.44	5.66	0.48
50	7.11	**13.88**	**9.96**	2.03	4.47	5.97	5.37	0.49	4.71	5.98	5.51	0.40
60	**7.66**	**15.04**	**9.96**	2.17	4.65	5.87	5.35	0.41	4.88	5.88	5.46	0.33
70	**7.66**	**14.45**	**9.99**	2.05	4.98	6.00	5.41	0.32	5.02	6.00	5.50	0.29
80	7.46	**14.30**	**9.76**	2.06	4.92	5.63	5.28	0.26	5.00	5.70	5.35	0.23
90	7.38	**14.25**	**9.81**	2.01	4.84	5.74	5.29	0.27	5.04	5.76	5.34	0.23
100	**7.53**	**14.11**	**9.81**	1.92	4.73	5.73	5.26	0.31	4.98	5.80	5.33	0.26
120	**7.88**	**14.11**	**9.78**	1.85	4.82	5.51	5.22	0.21	4.94	5.51	5.27	0.18
150	7.02	**13.99**	**9.60**	2.06	4.50	5.59	5.08	0.31	4.56	5.63	5.11	0.30
180	7.32	**14.06**	**9.53**	1.97	4.67	5.39	5.09	0.23	4.73	5.39	5.12	0.23
210	7.07	**13.97**	**9.50**	2.05	4.80	5.37	5.02	0.20	4.83	5.37	5.06	0.19
240	7.20	**13.95**	**9.63**	2.07	4.69	5.64	5.22	0.28	4.71	5.68	5.25	0.27
270	7.46	**13.74**	**9.42**	1.95	4.75	5.43	5.06	0.20	4.78	5.43	5.08	0.20
300	**7.80**	**13.77**	**9.66**	1.86	4.37	5.48	5.11	0.31	4.37	5.48	5.13	0.32
Total	7.02	**17.31**	**10.04**	2.11	3.88	**7.88**	5.43	0.66	4.37	**8.73**	5.61	0.72

Table 5

Type I error rates (in percentages) for the F-statistic, F-GG, and F-HF by N with distribution 9 ($\gamma_1 = 1.43, \gamma_2 = 3$) for all K (3, 4, and 6) and $\hat{\varepsilon} = .70$ and .80. Type I error rates > 7.5 are in bold (liberal).

N	F				F-GG				F-HF			
	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*
10	5.91	**7.81**	6.65	0.79	2.88	4.54	3.81	0.73	4.66	5.73	5.02	0.37
15	5.42	7.26	6.48	0.69	3.10	4.81	4.06	0.61	4.51	5.60	4.90	0.39
20	5.90	7.46	6.44	0.57	3.57	5.21	4.33	0.56	4.50	5.80	4.93	0.47
25	5.63	7.40	6.78	0.68	3.92	5.39	4.62	0.64	4.54	5.85	5.17	0.48
30	5.95	7.14	6.64	0.46	4.02	5.02	4.55	0.46	4.48	5.48	4.99	0.39
40	5.81	7.32	6.52	0.64	4.23	5.15	4.61	0.32	4.53	5.39	4.94	0.32
50	5.73	6.83	6.28	0.38	3.82	5.01	4.43	0.38	4.26	5.16	4.71	0.31
60	5.75	6.78	6.39	0.37	4.17	5.07	4.70	0.36	4.37	5.21	4.91	0.33
70	6.20	7.24	6.59	0.37	4.43	5.09	4.80	0.24	4.78	5.14	4.98	0.14
80	6.01	**7.61**	6.67	0.59	4.66	5.05	4.90	0.14	4.81	5.37	5.08	0.23
90	5.77	7.43	6.57	0.59	4.63	5.14	4.87	0.21	4.77	5.20	4.99	0.17
100	5.61	6.73	6.42	0.43	4.42	5.13	4.77	0.28	4.57	5.25	4.88	0.27
120	6.28	6.78	6.59	0.20	4.50	5.22	4.95	0.26	4.75	5.28	5.07	0.22
150	5.97	7.26	6.59	0.55	4.50	5.39	4.94	0.35	4.66	5.39	5.03	0.30
180	5.55	**7.57**	6.46	0.70	4.09	5.31	4.84	0.42	4.18	5.39	4.91	0.43
210	5.96	7.28	6.53	0.45	4.76	5.12	4.98	0.14	4.84	5.16	5.04	0.12
240	5.83	7.18	6.41	0.51	4.59	5.16	4.80	0.20	4.63	5.16	4.84	0.19
270	6.16	**7.68**	6.67	0.58	4.65	5.28	5.04	0.23	4.69	5.36	5.10	0.24
300	5.38	7.06	6.28	0.57	4.43	5.08	4.73	0.21	4.46	5.09	4.77	0.21
Total	5.38	**7.81**	6.52	0.52	2.88	5.39	4.67	0.48	4.18	5.85	4.96	0.31

Table 6

Type I error rates (in percentages) for the F-statistic, F-GG, and F-HF by N across distributions 10-11 (γ_1 = 2, γ_2 = 6; γ_1 = 2.31, γ_2 = 8) for all K (3, 4, and 6) and $\hat{\varepsilon} \leq .60$. Type I error rates > 7.5 are in bold (liberal).

N	*F*				*F-GG*				*F-HF*			
	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*
10	**8.56**	**19.76**	**13.40**	3.33	3.44	**11.81**	**8.19**	2.51	5.25	**12.32**	**9.17**	2.07
15	**8.34**	**18.58**	**12.50**	2.82	4.02	**10.07**	**7.78**	1.87	5.13	**10.20**	**8.33**	1.61
20	**8.46**	**17.37**	**11.75**	2.59	4.26	**9.26**	7.17	1.49	5.18	**9.33**	**7.56**	1.27
25	**8.48**	**16.58**	**11.40**	2.34	4.68	**8.55**	6.87	1.22	5.20	**8.59**	7.17	1.05
30	**8.50**	**16.30**	**11.16**	2.21	4.63	**7.92**	6.62	1.04	5.33	**8.04**	6.87	0.86
40	**7.94**	**15.35**	**10.72**	2.22	4.74	7.50	6.30	0.85	5.05	7.50	6.48	0.75
50	**7.87**	**15.43**	**10.47**	2.12	4.33	7.31	6.09	0.80	4.71	7.32	6.22	0.73
60	**7.88**	**15.31**	**10.38**	2.13	4.76	7.06	5.98	0.70	4.98	7.17	6.09	0.64
70	**7.63**	**15.13**	**10.27**	2.19	4.80	6.78	5.78	0.58	5.09	6.78	5.88	0.51
80	**7.97**	**14.62**	**10.28**	1.98	4.89	6.75	5.82	0.47	5.08	6.75	5.90	0.41
90	**7.80**	**14.88**	**10.20**	1.96	4.84	6.26	5.70	0.39	4.99	6.40	5.76	0.36
100	**7.87**	**14.23**	**10.09**	1.96	4.67	6.30	5.61	0.42	4.84	6.37	5.68	0.38
120	7.35	**14.03**	**9.92**	1.87	4.63	6.31	5.44	0.35	4.80	6.31	5.51	0.32
150	7.26	**14.54**	**9.91**	1.94	4.72	6.27	5.39	0.37	4.72	6.27	5.43	0.36
180	7.44	**14.03**	**9.81**	1.81	4.77	5.94	5.35	0.35	4.85	5.94	5.38	0.33
210	7.39	**13.71**	**9.66**	1.79	4.57	5.58	5.25	0.29	4.64	5.62	5.28	0.27
240	7.24	**13.82**	**9.65**	1.83	4.76	5.63	5.23	0.25	4.84	5.64	5.26	0.23
270	7.38	**14.24**	**9.75**	2.02	4.72	5.98	5.24	0.36	4.75	5.98	5.26	0.35
300	**7.56**	**13.83**	**9.70**	1.80	4.43	5.76	5.21	0.33	4.52	5.76	5.24	0.32
Total	7.24	**19.76**	**10.58**	2.37	3.44	**11.81**	6.05	1.29	4.52	**12.32**	6.23	1.36

Table 7

Type I error rates (in percentages) for the F-statistic, F-GG, and F-HF by N across distributions 10-11 (γ_1 = 2, γ_2 = 6; γ_1 = 2.31, γ_2 = 8) for all K (3, 4, and 6) and $\hat{\varepsilon}$ = .70 and .80. Type I error rates > 7.5 are in bold (liberal), those < 2.5 are in italics (conservative).

N	*F*				*F-GG*				*F-HF*			
	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*
10	4.97	**7.99**	6.46	1.00	*2.18*	4.93	3.33	0.86	3.60	6.19	4.50	0.76
15	5.34	**7.52**	6.41	0.78	2.52	5.17	3.77	0.80	3.82	5.87	4.56	0.63
20	5.44	**7.62**	6.45	0.72	3.12	5.26	4.03	0.68	4.01	5.83	4.60	0.58
25	5.76	**7.66**	6.56	0.68	3.49	5.58	4.25	0.67	4.15	6.01	4.75	0.60
30	5.42	**7.54**	6.54	0.64	3.66	5.26	4.37	0.56	4.30	5.77	4.79	0.48
40	5.87	7.29	6.46	0.43	3.82	5.44	4.43	0.54	4.32	5.78	4.77	0.46
50	6.02	7.41	6.65	0.54	4.09	5.52	4.63	0.45	4.39	5.80	4.89	0.41
60	5.63	7.42	6.45	0.57	4.00	5.15	4.62	0.33	4.36	5.31	4.84	0.27
70	5.70	7.45	6.60	0.58	4.34	5.20	4.76	0.29	4.57	5.31	4.95	0.24
80	5.44	7.35	6.47	0.60	4.34	5.36	4.68	0.31	4.45	5.48	4.84	0.32
90	5.91	7.11	6.45	0.39	4.30	5.10	4.72	0.28	4.45	5.23	4.86	0.24
100	5.64	6.97	6.41	0.45	4.11	5.11	4.60	0.33	4.35	5.24	4.74	0.27
120	5.93	7.45	6.51	0.53	4.24	5.13	4.74	0.26	4.42	5.18	4.85	0.25
150	5.65	7.44	6.44	0.58	4.36	5.24	4.74	0.25	4.40	5.30	4.83	0.25
180	5.71	7.37	6.49	0.53	4.47	5.37	4.87	0.31	4.57	5.40	4.94	0.30
210	5.60	6.81	6.37	0.38	4.20	5.01	4.72	0.27	4.34	5.10	4.79	0.25
240	5.76	7.29	6.58	0.43	4.57	5.37	4.91	0.27	4.65	5.43	4.96	0.27
270	5.51	**7.80**	6.60	0.59	4.50	5.40	4.97	0.24	4.54	5.50	5.02	0.24
300	5.40	**7.66**	6.57	0.56	4.50	5.26	4.89	0.24	4.52	5.33	4.93	0.25
Total	4.97	**7.99**	6.50	0.58	*2.18*	5.58	4.53	0.61	3.60	6.19	4.81	0.41

Table 8

Type I error rates (in percentages) for the F-statistic, F-GG, and F-HF by N across distributions 1-11 for all K (3, 4, and 6) and $\hat{\varepsilon}$ = .90. Type I error rates > 7.5 are in bold (liberal), those < 2.5 are in italics (conservative).

N	F				F-GG				F-HF			
	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*
10	4.75	6.04	5.41	0.34	*1.45*	4.84	3.18	0.89	3.00	5.42	4.41	0.55
15	4.94	6.22	5.59	0.33	*2.02*	4.94	3.71	0.79	3.17	5.38	4.68	0.58
20	5.10	6.38	5.50	0.32	*2.39*	5.34	3.90	0.67	3.30	5.66	4.64	0.46
25	5.00	6.19	5.54	0.27	2.86	5.26	4.14	0.67	3.59	5.56	4.74	0.47
30	5.04	6.34	5.65	0.32	2.89	5.00	4.31	0.54	3.67	5.52	4.84	0.39
40	4.48	6.30	5.56	0.36	3.24	5.26	4.37	0.45	3.82	5.44	4.80	0.37
50	4.80	6.26	5.57	0.32	3.52	5.52	4.52	0.46	4.10	5.76	4.87	0.40
60	4.74	6.42	5.63	0.39	3.60	5.52	4.60	0.39	4.00	5.64	4.91	0.38
70	5.18	6.02	5.53	0.20	3.78	5.06	4.59	0.30	4.18	5.40	4.83	0.24
80	4.66	6.30	5.52	0.30	3.88	5.20	4.64	0.32	4.16	5.54	4.85	0.30
90	5.14	6.60	5.61	0.36	4.10	5.48	4.70	0.35	4.38	5.70	4.90	0.33
100	4.90	6.04	5.52	0.27	4.07	5.14	4.64	0.26	4.20	5.28	4.84	0.26
120	4.90	5.98	5.62	0.24	4.20	5.28	4.81	0.25	4.38	5.38	4.96	0.23
150	5.06	6.06	5.52	0.29	4.19	5.40	4.80	0.31	4.33	5.46	4.91	0.30
180	4.58	6.68	5.54	0.41	3.98	5.78	4.79	0.37	4.10	5.90	4.91	0.37
210	4.78	6.28	5.52	0.33	4.14	5.32	4.80	0.32	4.24	5.54	4.88	0.32
240	4.96	6.04	5.55	0.27	4.32	5.42	4.86	0.25	4.36	5.44	4.94	0.24
270	5.08	6.32	5.64	0.31	4.46	5.52	4.97	0.29	4.52	5.60	5.03	0.30
300	5.16	6.26	5.62	0.27	4.52	5.52	4.95	0.26	4.63	5.60	5.02	0.25
Total	4.48	6.68	5.56	0.32	*1.45*	5.78	4.49	0.65	3.00	5.90	4.84	0.39

Considering Bradley's (1978) liberal criterion, and with distributions 1-8 (with γ_1 up to 1 and γ_2 up to 3), the *F*-statistic is liberal with $\hat{\varepsilon} \leq .60$ and, in some conditions, with $\hat{\varepsilon} = .70$ and .80. *F-HF* and *F-GG* are robust in all cases.

With distribution 9 (with $\gamma_1 = 1.41$ and $\gamma_2 = 3$), the *F*-statistic is liberal with $\hat{\varepsilon} \leq .60$, but *F-HF* and *F-GG* are generally robust, except with small sample size ($N = 10$). With $\hat{\varepsilon} = .70$ and .80, the *F*-statistic is liberal in some cases, whereas *F-GG* and *F-HF* are robust.

With distributions 10-11 (with $\gamma_1 = 2$ or 2.31 and $\gamma_2 = 6$ or 8), the *F*-statistic is liberal with $\hat{\varepsilon} \leq .60$, and *F-HF* and *F-GG* are also liberal with sample sizes equal to or below 30. With $\hat{\varepsilon} = .70$ and .80, the *F*-statistic is liberal in some cases, *F-GG* can become conservative with $N = 10$, and *F-HF* is robust in all conditions.

With $\hat{\varepsilon} = .90$, and for all *K* and distributions, the *F*-statistic was within the interval [2.5, 7.5] for considering a test as robust. *F-GG* can become conservative for *N* as small as 10, but *F-HF* is robust under all conditions.

It can be seen in Table 9, which displays results according to Bradley's (1978) stringent criterion, that *F-GG* tends to be more conservative than *F-HF*, and also that the percentage robustness of *F-GG* is greater than that of *F-HF* with $\hat{\varepsilon} < .60$, and lower with $\hat{\varepsilon} \geq .60$. With $\hat{\varepsilon} = .90$, *F-HF* outperforms the *F*-statistic in terms of percentage robustness.

Table 9

Percentage robustness of the F-statistic, F-GG, and F-HF according to Bradley's stringent criterion. Conservative: Type I error < 4.5; robust: falls in the interval [4.5, 5.5]; liberal: > 5.5. Shaded boxes indicate higher percentage robustness for each value of $\hat{\varepsilon}$.

	F			F-GG			F-HF		
$\hat{\varepsilon}$	Conservative	Robust	Liberal	Conservative	Robust	Liberal	Conservative	Robust	Liberal
.30[a]	-	-	100	2.4	64.6	33.0	2.4	56.5	41.1
.40[b]	-	-	100	2.4	66.0	31.6	1.4	57.9	40.7
.50	-	-	100	10.3	75.1	14.6	1.7	72.7	25.6
.60	-	-	100	13.4	73.4	13.2	2.9	75.8	21.4
.70	-	-	100	23.1	75.9	1.0	6.9	84.4	8.8
.80	-	4.0	96.0	33.7	65.6	0.8	9.9	86.3	3.8
.90	0.2	43.4	56.5	35.9	63.3	0.8	14.2	83.1	2.7

Note. Given that the same estimated ε value of *F-GG* and *F-HF* is reached with the lower limit of $\hat{\varepsilon}$ for each *K*, these percentages have been eliminated from the computation. [a] For $K = 6$; [b] for $K = 4$ and 6.

Empirical Power

The empirical power of the *F*-statistic, *F-GG,* and *F-HF* showed the same behavior across mean patterns in each *K* as a function of distribution shape, sphericity, and *N*. Figures 1-3 display empirical power for the three statistics with these variables collapsed by mean patterns. Table 10 shows the *N* at which a power of 80% is reached in each manipulated condition. We have removed the power of the *F*-statistic when $\hat{\varepsilon} \leq .60$ because it was liberal in all conditions. Detailed tables are available as supplementary material.

Figure 1

Percentage empirical power as a function of distribution shape, sphericity ($\hat{\varepsilon}$), and sample size for K = 3. In parenthesis: skewness (γ_1) and kurtosis (γ_2) coefficients.

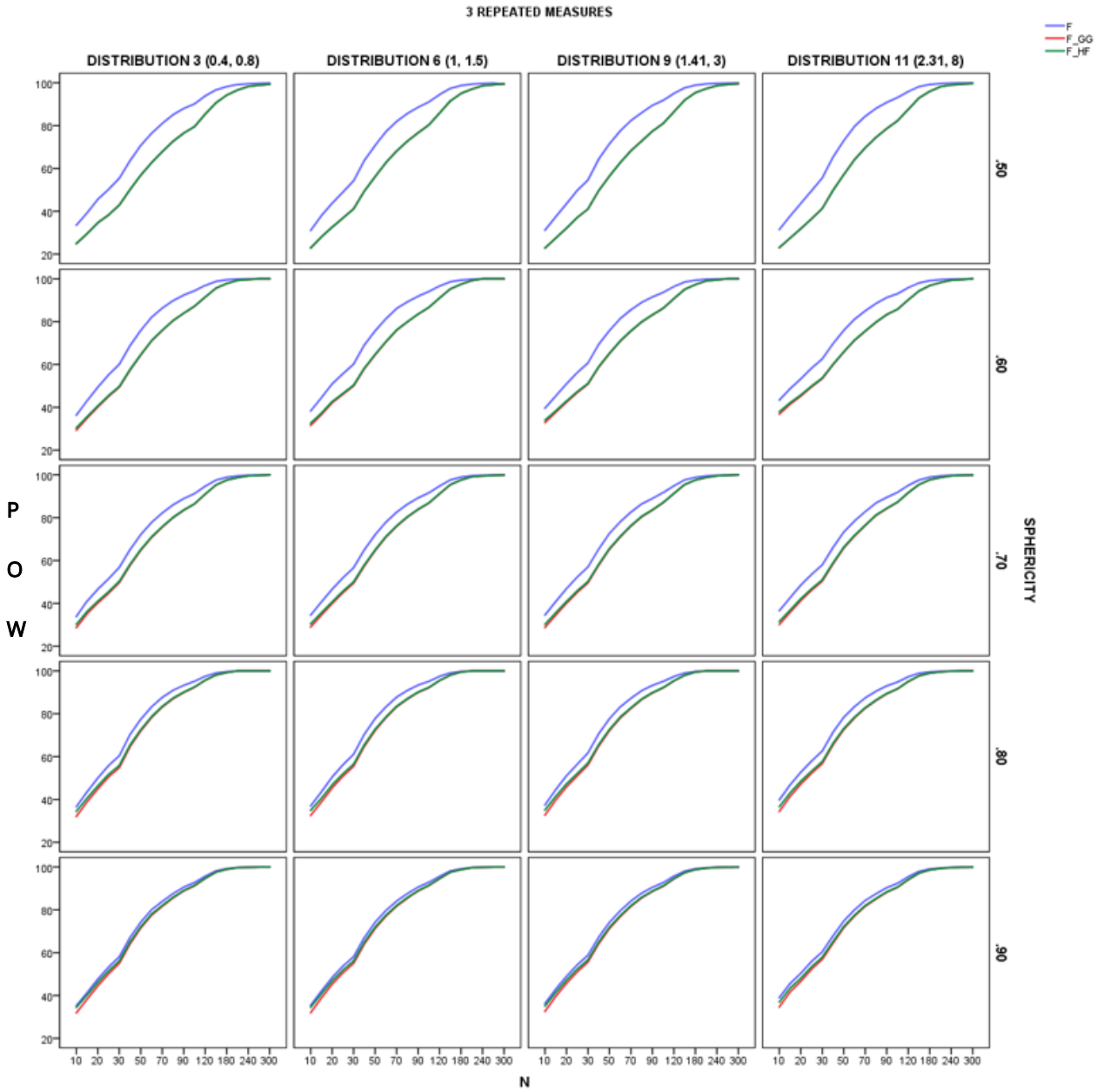


Figure 2

Percentage empirical power as a function of distribution shape, sphericity ($\tilde{\varepsilon}$), and sample size for K = 4. In parenthesis: skewness (γ_1) and kurtosis (γ_2) coefficients.

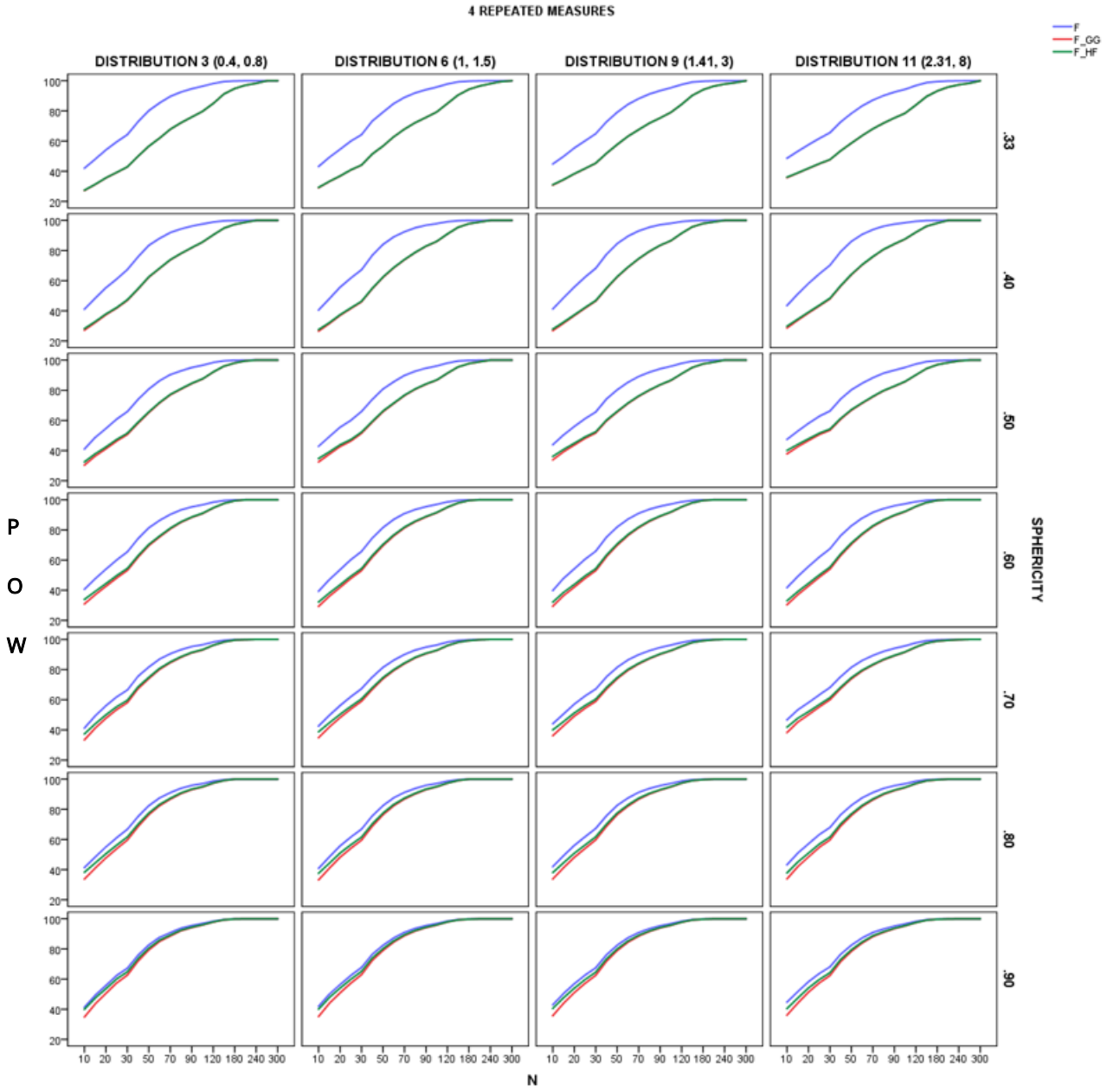


Figure 3

Percentage empirical power as a function of distribution shape, sphericity ($\hat{\varepsilon}$), and sample size for K = 6. In parenthesis: skewness (γ_1) and kurtosis (γ_2) coefficients.

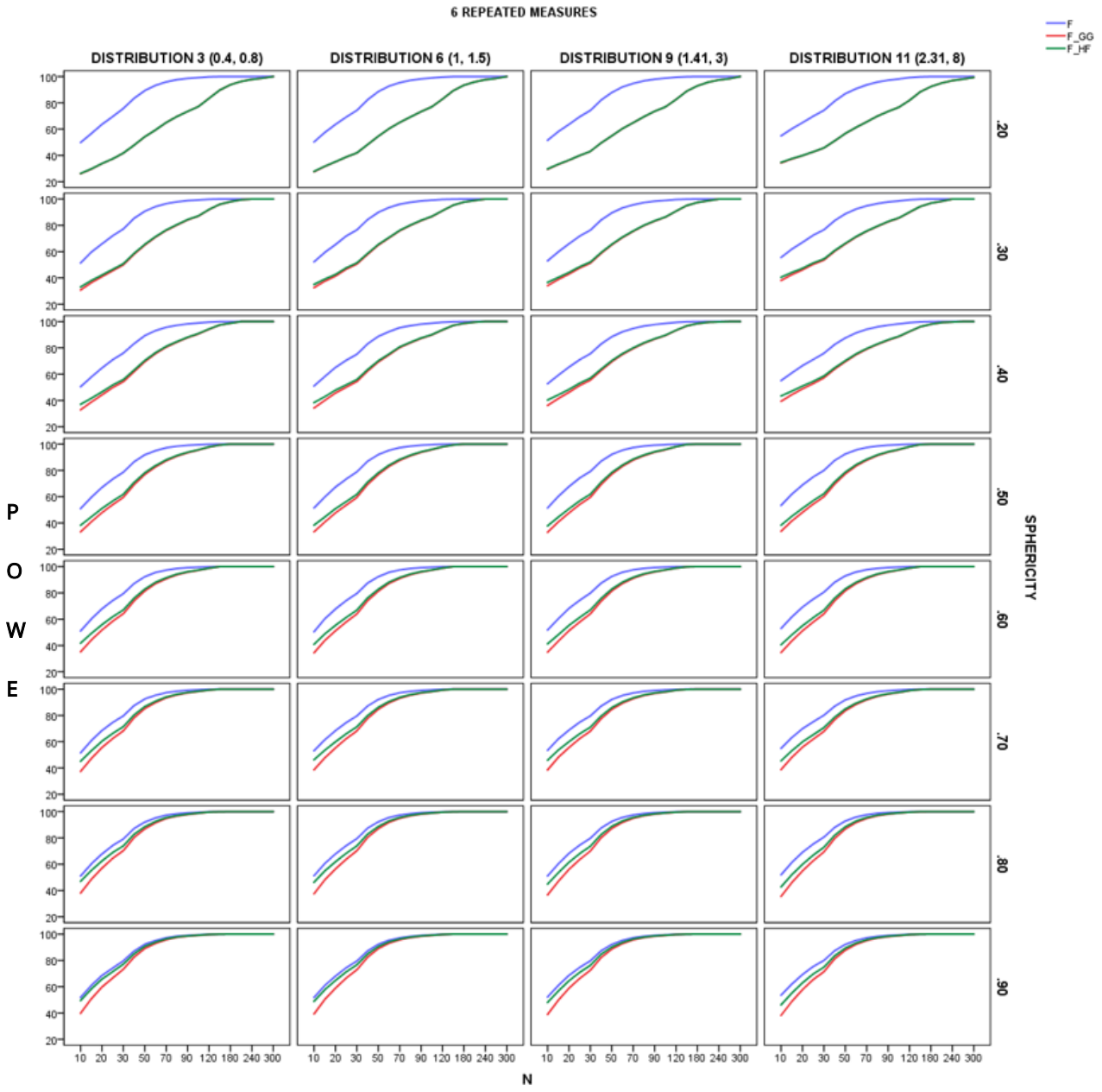


Table 10
Sample size at which a mean power of 80% is reached as a function of distribution shape, sphericity ($\hat{\varepsilon}$), and number of repeated measures (K) across all mean patterns. In parenthesis: skewness (γ_1) and kurtosis (γ_2) coefficients.

		Distribution 3 (0.4, 0.8)			Distribution 6 (1, 1.5)			Distribution 9 (1.41, 3)			Distribution 11 (2.31, 8)		
K	$\hat{\varepsilon}$	*F*	*F-GG*	*F-HF*	*F*	*F-GG*	*F-HF*	*F*	*F-GG*	*F-HF*	*F*	*F-GG*	*F-HF*
3	.50	-	100	100	-	100	100	-	100	100	-	100	100
	.60	-	80	80	-	80	80	-	80	80	-	80	80
	.70	70	80	80	70	80	80	70	80	80	70	80	80
	.80	60	70	70	60	70	70	60	70	70	60	70	70
	.90	60	70	70	60	70	70	60	70	70	60	70	70
4	.33	-	100	100	-	120	120	-	120	120	-	120	120
	.40	-	90	90	-	90	90	-	90	80	-	80	80
	.50	-	80	80	-	80	80	-	80	80	-	80	80
	.60	-	70	70	-	70	70	-	70	70	-	70	70
	.70	50	60	60	50	70	60	50	60	60	50	70	60
	.80	50	60	60	50	60	60	50	60	60	50	60	60
	.90	50	60	50	50	60	50	50	60	50	50	60	60
6	.20	-	120	120	-	120	120	-	120	120	-	120	120
	.30	-	80	80	-	80	80	-	80	80	-	90	90
	.40	-	70	70	-	70	70	-	80	70	-	80	70
	.50	-	60	60	-	60	60	-	60	60	-	60	60
	.60	-	50	50	-	50	50	-	50	50	-	50	50
	.70	30	50	40	30	50	40	30	50	40	30	50	50
	.80	40	40	40	40	40	40	30	40	40	30	40	40
	.90	30	40	40	30	40	40	30	40	40	30	40	40

Overall, power increases as sample size increases, the *F*-statistic shows greater power than do the two adjusted *F*-tests, and the power of *F-HF* is slightly greater than that of *F-GG* for small sample size as sphericity increases, especially for $K = 4$ and 6. The same profile of power for the three statistics is observed across distributions for each *K*.

Discussion

The purpose of this study was to conduct a detailed examination of Type I error and power of the *F*-statistic, *F-GG,* and *F-HF* under a wide number of conditions involving simultaneous violation of normality and sphericity, as may be encountered in real research situations. Our ultimate goal was to clarify the conditions in which each procedure may be used. To this end, we manipulated the number of repeated measures ($K = 3$, 4, and 6), sample size (from 10 to 300), sphericity ($\hat{\varepsilon}$, from its lower limit to .90, as a function of *K*), and shape of the distribution, from slight to extreme deviations from normality.

Overall, the results show that Type I error rates of the *F*-statistic, *F-GG*, and *F-HF* depend on the degree of deviation from the normal distribution, the degree of sphericity violation, and sample size.

Considering Bradley's (1978) liberal criterion, the results for distributions with $\gamma_1 \leq 1$ and $\gamma_2 \leq 3$ indicate that the *F*-statistic tends to be liberal under violation of sphericity. *F-GG* and *F-HF* are robust in all conditions and are closer to 5%. Therefore, in the presence of non-normal data with the above values of skewness and kurtosis, both *F-GG* and *F-HF* can be used with violations of sphericity while still ensuring that Type I error is in the interval [2.5, 7.5].

For a distribution with $\gamma_1 = 1.41$ and $\gamma_2 = 3$, the *F*-statistic shows approximately the same behavior as with the aforementioned distributions, although its tendency to be liberal increases. Overall, *F-GG* and *F-HF* are robust, but with $\hat{\varepsilon} \leq .60$ and a sample size as small as 10, their Type I error can become inflated. These results suggest that with severe deviations from normality and sphericity, and very small sample size, these adjusted-*F* tests should be avoided.

With distributions representing extreme deviation from normality, with $\gamma_1 \approx 2$ and $6 \leq \gamma_2 \leq 8$, the tendency of the *F*-statistic to be liberal is exacerbated. The Type I error of *F-GG* and *F-HF* depends on the $\hat{\varepsilon}$ value and sample size. With $\hat{\varepsilon} \leq .60$, both these adjusted tests tend to be liberal with sample size equal to or less than 30, and robust with larger sample sizes. With $\hat{\varepsilon} = .70$ and .80, *F-GG* can become conservative with $N = 10$, whereas *F-HF* is robust in all conditions

In all distributions, when $\hat{\varepsilon} = .90$ the Type I error of the *F*-statistic and *F-HF* are robust, whereas *F-GG* can become conservative for *N* as small as 10.

When applying Bradley's (1978) stringent criterion of robustness to achieve a more refined analysis, the results show that although *F-GG* tends to be more conservative than *F-HF*, the robustness of both procedures depends on the $\hat{\varepsilon}$ value: *F-GG* is superior to *F-HF* with $\hat{\varepsilon} < .60$, and *F-HF* is superior to *F-GG* with $\hat{\varepsilon} \geq .60$. In addition, *F-HF* is superior to the *F*-statistic for large values of $\hat{\varepsilon}$, even when $\hat{\varepsilon} = .90$.

Regarding empirical power, the results show that power increases as sample size increases, and also that the *F*-

statistic shows greater power than either of the two adjusted *F*-tests. These results are expected as they reflect the known relationship between power and sample size, and between power and Type I error. The power of *F-HF* is slightly greater than that of *F-GG* as sphericity increases for small sample size, especially for designs with a higher number of repeated measures. The same profile of power for the three statistics is observed across distributions, values of $\hat{\varepsilon}$, and number of repeated measures. Power decreases with lower values of $\hat{\varepsilon}$, which indicates that a larger sample size is needed to reach a power of 80% for a medium effect size. For example, for $K = 3$ and $\hat{\varepsilon} = .50$ the sample size required is 100, whereas for $\hat{\varepsilon} = .90$ it is 60.

These results highlight the following issues:

1. The *F*-statistic is liberal with violation of sphericity. The more severe the violation, the more liberal it is. This result has been consistently found in previous research (Berkovits et al., 2000; Blanca et al., 2023b; Box, 1954; Collier et al., 1967; Haverkamp & Beauducel, 2017, 2019; Voelkle & McKnight, 2012).
2. The tendency toward liberality of the *F*-statistic with violation of sphericity is aggravated with severe violation of normality. Blanca et al. (2023a) found that non-normality does not affect robustness of the *F*-statistic when sphericity is met. Our finding here therefore extends knowledge, showing that severe non-normality does have an impact on robustness when sphericity is simultaneously violated.
3. Overall, *F-GG* tends to be more conservative than *F-HF*. This has been reported previously (Blanca et al., 2023b; Haverkamp & Beauducel, 2017; Huynh & Feltdt, 1976; Oberfeld & Franke, 2013) and has led some authors to recommend, as a general rule, the use of *F-GG* over *F-HF* (Kirk, 2013; Maxwell & Delaney, 2004; Voelkle & McKnight, 2012).
4. Violation of normality and sphericity has an impact on the robustness of *F-GG* and *F-HF* with small sample size ($N \leq 30$), and both statistics tend to be liberal with severe violation of both normality and sphericity ($\hat{\varepsilon} \leq .60$). Berkovits et al. (2000) obtained similar results, but as they only considered four sample sizes (10, 15, 30, and 60), it was not possible to determine more precisely the sample size at which the change from liberality to robustness of these statistics occurred.
5. Application of Bradley's (1978) stringent criterion of robustness indicates that *F-GG* outperforms *F-HF* with $\hat{\varepsilon} < .60$, while *F-HF* outperforms *F-GG* with $\hat{\varepsilon} \geq .60$. This can help to establish guidelines for RM-ANOVA in the event of discrepant results from these two statistics. Our findings here are in line with Blanca et al. (2023b) and establish a more restrictive cut-off for the use of *F-GG* and *F-HF* than has been proposed previously. For example, Huynh and Feldt (1976) and Barcikowski and Robey (1984) set the threshold at .75.
6. *F-HF* is slightly more powerful than *F-GG* for larger $\hat{\varepsilon}$ values with small sample size, although the two have equivalent power with large samples. This finding has been reported previously (Algina & Keselman, 1997; Blanca et al., 2023b) and may be explained by the tendency of *F-GG* to be more conservative than *F-HF*.
7. Overall, the more severe the sphericity violation, the larger the sample size needed to achieve 80% power for a medium effect size. It is important to take this into consideration when planning research.

Practical recommendations

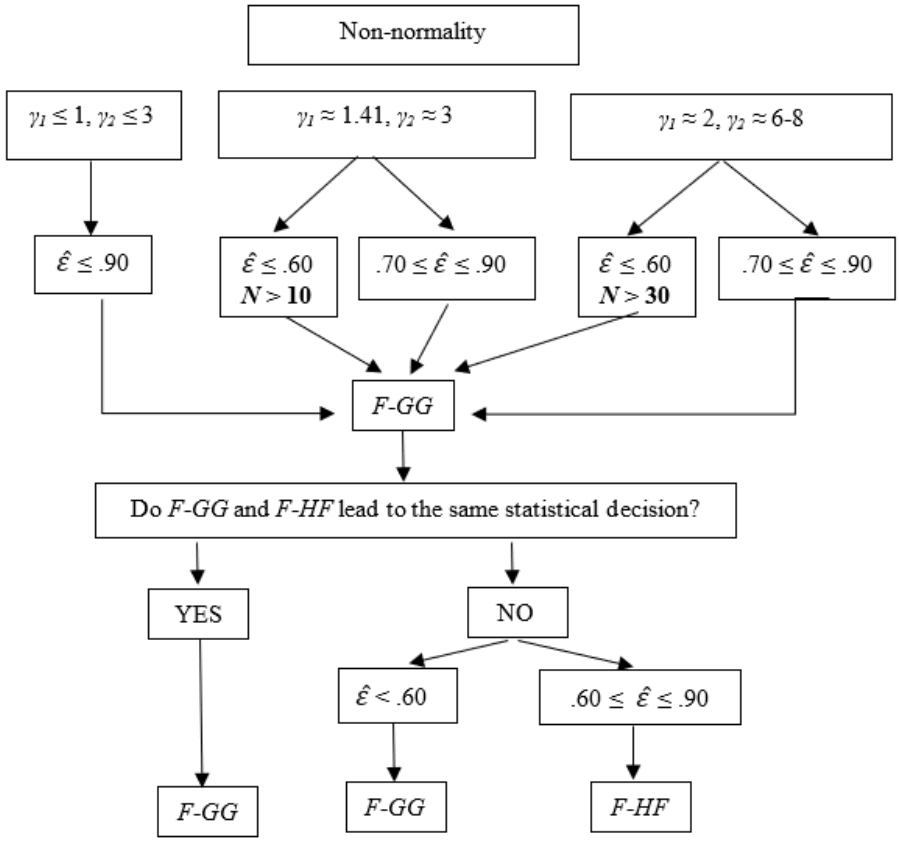
A number of practical recommendations may be proposed based on the results. First, in order to keep Type I error within the interval [2.5, 7.5] when conducting RM-ANOVA, researchers should consider three key aspects: degree of deviation from normality, degree of sphericity violation, and sample size. Although both *F-GG* and *F-HF* may be adequate alternatives to the *F*-statistic in some conditions, our recommendation, in the event that the two adjusted *F*-tests lead to the same statistical decision, is to use and report *F-GG* as it shows more conservative behavior than does *F-HF*. The former may be used under violation of sphericity and slight or moderate deviations from normality, that is, with asymmetry and kurtosis coefficients equal to or lower than 1 and 3, respectively. With severe deviations from normality, for example, with asymmetry and kurtosis coefficients around 1.40 and 3, *F-GG* may be used with $\hat{\varepsilon} \geq .70$ but with $\hat{\varepsilon} \leq .60$ a sample size larger than 10 is required. With extreme deviation from normality (asymmetry and kurtosis coefficients around 2 and 6-8), this statistic may be used with $\hat{\varepsilon} \geq .70$, and with a sample size larger than 30 for $\hat{\varepsilon} \leq .60$.

As a general rule, therefore, *F-GG* is a suitable alternative to the *F*-statistic when the data are non-normally distributed and sphericity is violated, provided that the sample size is larger than 30. The greater the deviation from normality (high values of asymmetry and kurtosis coefficients) and the violation of sphericity (lower values of $\hat{\varepsilon}$), the larger the sample size required to ensure the robustness and adequate power of these procedures. A power of 80% is usually used when a priori analysis of sample size is performed (Cooper & Garson, 2016; Kirk, 2013). We encourage researchers to perform this a priori power analysis to estimate the sample size required, considering potential distributional characteristics with an approximate expected value of sphericity. G*Power software may be especially useful for this purpose (Faul et al., 2007).

There is also the question of what to do when the *F-GG* and *F-HF* procedures yield discrepant results. For example, *F-GG* leads the researcher to accept the null hypothesis of mean differences, whereas according to *F-HF* it should be rejected. In such situations, and taking the Greenhouse-Geisser epsilon estimate as a reference, we recommend the

use of *F-GG* with $\hat{\varepsilon} < .60$ and *F-HF* with $\hat{\varepsilon} \geq .60$. *F-HF* should be used even with $\hat{\varepsilon} = .90$. This rule of thumb is in line with that proposed by Blanca et al. (2023b), who established it under violation of sphericity with normal data. The present results extend this rule to situations involving simultaneous violation of normality and sphericity. Figure 4 summarizes the analytic strategies that follow from these recommendations.

Figure 4
Analytic strategies as a function of the results of the simulation study (γ_1: skewness coefficient; γ_2: kurtosis coefficient)



It should be noted that these practical recommendations are only valuable under certain circumstances. Although they can guide a large number of real research situations, they do not provide a solution to scenarios in which severe violations of normality and sphericity coexist with samples equal to or less than 30 and in which the researcher is unable to increase the sample size. Several statistical alternatives to ANOVA and adjusted *F*-tests have been proposed, including classical non-parametric analysis (the Friedman test), multivariate analysis, the linear mixed model, and the bootstrap method. However, the results of simulation studies suggest that the behavior of these statistical procedures is far from clear under the circumstances mentioned above. For example, with small samples the Friedman test has been found to be robust when sphericity is violated with normal data (Harwell & Serlin, 1994; Hayoz, 2007), and also for some non-normal distributions but with spherical data (Al-Subaihi, 2000). Multivariate analysis has been shown to be robust with $N = 25$ for 4 and 6 repeated measures and an epsilon value of .50 (Voelke & McKnight, 2012), although other studies have observed a tendency toward liberality with $N < 30$ and severe violations of sphericity and normality, under which conditions this approach performs worse than *F-GG* and *F-HF* (Berkovits et al., 2000). The linear mixed model (LMM), which does not require fulfillment of a strict sphericity assumption, although it can account for different covariance structures (Muhammad, 2023), has been found to perform worse than *F-HF* when the sphericity assumption is violated, the sample size is quite small, and the number of repeated measures is large (Haverkamp & Beauducel, 2017). The results of other studies also suggest that use of the Kenward-Roger correction with the LMM does not control type I error when $N < 30$ (Haverkamp & Beauducel, 2019). These divergent results are probably due to differences in the conditions manipulated in simulation studies, but overall they suggest that none of these procedures can reliably be considered adequate under conditions of non-normality and non-sphericity with samples equal to or less than 30. Further research is warranted to clarify the behavior of these procedures under these scenarios.

The most promising alternative in those scenarios where adjusted *F*-tests do not provide valid results may be the bootstrap method. Berkovits et al. (2000) found that bootstrap-*F* appeared to offer reasonable Type I error control under violation of both normality and sphericity, even with fairly small sample size. However, these authors only analyzed Type I error in a limited number of conditions, namely four non-normal distributions, sample size equal to or less than 60, and values of epsilon of .48, .57, and .75 for a one-way design with four repeated measures. Further research is needed to deepen and extend knowledge of the behavior of this technique, examining both Type I error and power and increasing the number of conditions manipulated.

Complementary information

Acknowledgements.- The authors would like to thank Macarena Torrado for her collaboration in this study.
Funding.- This research was supported by grant PID2020-113191GB-I00, awarded through MCIN/AEI/10.13039/501100011033.
Conflict of interest.- The authors declare they have no conflict of interest or competing interests.

References

Al-Subaihi, A. A. (2000). A Monte Carlo study of the Friedman and Conover tests in the single-factor repeated measures design. *Journal of Statistical Computation and Simulation*, *65*(1-4), 203-223. https://doi.org/10.1080/00949650008811999

Armstrong, R. (2017). Recommendations for analysis of repeated-measures designs: Testing and correcting for sphericity and use of MANOVA and mixed model analysis. *Ophthalmic & Physiological Optics*, *37*(5), 585–593. https://doi.org/1.1111/opo.12399.

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*, *44*(4), 1224–1238. https://doi.org/10.3758/s13428-012-0196-y

Arnau, J., Bendayan, R., Blanca, M. J., & Bono, R. (2013). The effect of skewness and kurtosis on the robustness of linear mixed models. *Behavior Research Methods, 45*(3), 873–879. https://doi.org/10.3758/s13428-012-0306-x

Algina, J., & Keselman, H. (1997). Detecting repeated measures effects with univariate and multivariate statistics. *Psychological Methods*, *2*(2), 208–218. https://doi.org/10.1037/1082-989X.2.2.208

Barcikowski, R. S., & Robey, R. R. (1984). Decisions in single group repeated measures analysis: Statistical tests and three computer packages. *The American Statistician, 38*(2), 148–150.

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., Alarcón, R., & Bono, R. (2018). Current practices in data analysis procedures in psychology: What has changed? *Frontiers in Psychology*, *9*, Article 2558. https://doi.org/10.3389/fpsyg.2018.02558

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., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 9*(2), 78–84. https://doi.org/10.1027/1614-2241/a000057

Bono, R., Blanca, M. J., Arnau, J., & Gómez-Benito, J. (2017). Non-normal distributions commonly used in health, education, and social sciences: A systematic review. *Frontiers in Psychology, 8*, Article 1602. https://doi.org/10.3389/fpsyg.2017.01602

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*, 484–498. https://doi.org/10.1214/aoms/1177728717

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

Collier, R. O., Baker, F. B., Mandeville, G. K., & Hayes, T. F. (1967). Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures design. *Psychometrika, 32(3)*, 339–353. https://doi.org/10.1007/BF02289596

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

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–91. https://doi.org/10.3758/bf03193146

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*(4), 521–532. https://1.1007/BF02293811

Geisser, S. W., & Greenhouse, S. (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

Goedert, K., Boston, R., & Barrett, A. (2013). Advancing the science of spatial neglect rehabilitation: An improved statistical approach with mixed linear modeling. *Frontiers in Human Neuroscience*, *7*, Article 211. https://doi.org/10.3389/fnhum.2013.00211

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

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

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

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

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

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/1.1007/s00362-014-0617-3

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

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), 624–640. https://doi.org/10.1080/03610910903548952

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Lawrence Erlbaum Associates.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Muhammad, L. N. (2023). Guidelines for repeated measures statistical analysis approaches with basic science research considerations. *The Journal of Clinical Investigation, 133*(11), e171058. https://doi.org/10.1172/JCI171058

Muller, K. E., & Barton, C. N. (1989). Approximate power for repeated-measures ANOVA lacking sphericity. *Journal of the American Statistical Association, 84*(406), 549-555. https://doi.org/10.1080/01621459.1989.10478802

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

Sheskin, D. J. (2003). *Handbook of parametric and nonparametric statistical procedures*. Chapman and Hall/CRC.

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: European Journal of Research Methods for the Behavioral and Social Sciences, 8*, 23–38. https://doi.org/10.1027/1614-2241/a000044

Wilcox, R. R. (2022). *Introduction to robust estimation and hypothesis testing* (5th ed.). Academic Press.