

Evaluating the robustness of repeated measures analyses: The case of small sample sizes and nonnormal data

Daniel Oberfeld · Thomas Franke

Published online: 25 November 2012
© Psychonomic Society, Inc. 2012

Abstract Repeated measures analyses of variance are the method of choice in many studies from experimental psychology and the neurosciences. Data from these fields are often characterized by small sample sizes, high numbers of factor levels of the within-subjects factor(s), and nonnormally distributed response variables such as response times. For a design with a single within-subjects factor, we investigated Type I error control in univariate tests with corrected degrees of freedom, the multivariate approach, and a mixed-model (multilevel) approach (SAS PROC MIXED) with Kenward–Roger’s adjusted degrees of freedom. We simulated multivariate normal and nonnormal distributions with varied population variance–covariance structures (spherical and nonspherical), sample sizes (N), and numbers of factor levels (K). For normally distributed data, as expected, the univariate approach with Huynh–Feldt correction controlled the Type I error rate with only very few exceptions, even if sample sizes as low as three were combined with high numbers of factor levels. The multivariate approach also controlled the Type I error rate, but it requires $N \geq K$. PROC MIXED often showed acceptable control of the Type I error rate for normal data, but it also produced several liberal or conservative results. For nonnormal data, all of the procedures showed clear deviations from the nominal Type I error rate in many conditions, even for sample sizes greater than 50. Thus, none of these approaches can be considered robust if the response variable is nonnormally distributed. The results indicate that both the variance heterogeneity and covariance heterogeneity of the population covariance matrices affect the error rates.

Keywords Analysis of variance · Robustness · Nonnormality · Small sample settings · Repeated measurements · Correlated data · Multivariate · Mixed model analyses · Multilevel model · Simulation study · Type I error rate · Central limit theorem · Monte Carlo

Repeated measures designs in which each experimental unit (e.g., subject) is tested in more than one experimental condition are very common in psychology, the neurosciences, medicine, the social sciences, and agricultural research. The data from such experiments are often analyzed with analyses of variance (ANOVAs). The statistical assumptions that must be met in order for repeated measures ANOVAs to be valid are stronger than for data from a completely randomized design (also termed an *independent-groups design*) in which each experimental unit is tested only under one single experimental condition. For example, when testing for main effects and interactions with more than one numerator degree of freedom, the variance–covariance structure of the data is important for the validity of the tests (cf. Huynh & Feldt, 1970; Keselman, Algina, & Kowalchuk, 2001; Rouanet & Lépine, 1970). The extents to which several different approaches for the analysis of repeated measures data are robust, in the sense that they control the Type I error rate (i.e., the probability that the null hypothesis will be rejected, given that it is true), have been studied extensively for certain repeated measures designs (for reviews, see Keselman et al., 2001; Keselman, Algina, & Kowalchuk, 2002).

The aim of the present study was to investigate the control of the Type I error rate exhibited by different types of repeated measures ANOVAs in situations characterized by very small sample sizes (N as low as three) and nonnormality. Such data are frequently encountered in controlled experiments in the fields of experimental psychology, the neurosciences, and in certain clinical studies. However, the robustness of the different procedures for the analysis of this specific type of repeated

Electronic supplementary material The online version of this article (doi:10.3758/s13428-012-0281-2) contains supplementary material, which is available to authorized users.

D. Oberfeld (✉) · T. Franke
Department of Psychology, Johannes Gutenberg-Universität,
55099 Mainz, Germany
e-mail: oberfeld@uni-mainz.de

measures data has not yet been investigated very systematically. We illustrate three characteristics of the data type examined in our study with example studies from the relevant fields that have each been cited between 300 and 2,000 times, according to the Science Citation Index (<http://isiknowledge.com/>), and can thus be considered as relevant and accepted contributions to the respective fields of research. The three characteristics are (a) small or very small sample sizes (typically, $N = 3$ to 30), (b) nonnormally distributed dependent measures, and (c) the use of a completely within-subjects design (i.e., there are no between-subjects factors).

Characteristic (a) can be attributed to economic considerations. In many experiments—for example, in visual psychophysics or psychoacoustics—the experimentation time required for each subject is high. In these fields, typically large numbers of trials are collected per subject and experimental condition, because experimenters wish to minimize the effect of, for instance, day-to-day fluctuations on detection thresholds, or because they want to compute measures of sensitivity, such as d' from signal detection theory (cf. Green & Swets, 1966), that require several hundreds of trials. Additionally, researchers are often interested in the best possible performance that subjects can attain, in order to explore the limits of the sensory and cognitive systems. Therefore, highly trained subjects are required, and training periods of up to 5 h are not unusual (e.g., Duncan & Humphreys, 1989). As a consequence, to keep experimentation time at a manageable level, small numbers of subjects are tested ($N = 2$ to 10; e.g., Duncan & Humphreys, 1989; Eriksen & Eriksen, 1974; Ernst & Banks, 2002). If physiological responses are collected, there are additional reasons to restrict the number of subjects tested. For example, in electroencephalography (EEG) experiments, which nowadays often use up to 128 electrodes, the attachment of the electrodes is time consuming and also poses a certain risk of infection to the subjects, so that often small samples are used ($N = 5$ to 20; e.g., Sams, Paavilainen, Alho, & Näätänen, 1985). In neuroimaging studies using fMRI, operating the MRI scanner is expensive, not least because qualified medical and technical personnel are required. Additionally, the time available for research use is often very limited in clinics, due to the priority of medical use. Beyond that, the data analysis for a high number of subjects would be time consuming, because the fMRI data have to be mapped to anatomical structures on an individual basis (cf. Brett, Johnsrude, & Owen, 2002). Therefore, small sample sizes are again common ($N = 10$ to 20; e.g., Kanwisher, McDermott, & Chun, 1997). Other factors imposing limitations on the sample size are limited access to rare populations—such as, for example, persons with special types of synesthesia or patients with specific neurological disorders—or the ethical policy of minimizing the use of animals for research.

Concerning the distributional form of the response variables [characteristic (b)], it should be noted that not only in the research fields discussed here, but also in many other areas in psychology, strong deviations from normality are frequently encountered (Micceri, 1989). In psychophysics and cognitive neuroscience, two very important response measures are nonnormally distributed. *Response times* show a skewed and heavy-tailed distribution that is often successfully modeled as an ex-Gaussian distribution, which is a convolution of the Gaussian and exponential distributions (Luce, 1986; Van Zandt, 2000). In first approximation, response times follow a (shifted) log-normal distribution (Heathcote, Brown, & Cousineau, 2004; Ulrich & Miller, 1994). *Proportions* like error rates or the proportion of correct responses are binomially distributed. Other important measures, such as detection thresholds, d' values, or amplitudes of evoked EEG responses, however, can be considered to be normally distributed.

Finally, the use of designs containing only within-subjects factors [characteristic (c)] is due to the fact that, for example, in basic research on vision or audition, inter-individual differences are often considered less important because the aim is to understand the basic functioning of these sensory systems. Therefore, designs containing only within-subjects factors (i.e., each subject is tested under all experimental conditions) are advantageous because they provide comparably high power even with small sample sizes.

To investigate the robustness of repeated measures analyses for data showing the three characteristics discussed above, we simulated a design with a single within-subjects factor, no between-subjects factors, small sample sizes, and both normally and nonnormally distributed response variables. Due to the substantial computation time required for our simulations, we focused on Type I error rates (i.e., probability of false positives) and did not obtain Type II error rates (i.e., power; Algina & Keselman, 1997, 1998; Potvin & Schutz, 2000). In fundamental research, often more emphasis is placed on avoiding Type I errors than on avoiding Type II errors. Therefore, our data provide a basis for identifying procedures that control the Type I error rate in the specific situations that we studied. Additional simulation studies will be required to compare the procedures that we identified as suitable in terms of the Type I error rate with respect to statistical power (Type II error rate). Following many previous studies on the empirical Type I error rate for repeated measures designs, we adopted the “liberal” criterion of robustness proposed by Bradley (1978), according to which a procedure can be considered robust if the empirical Type I error rate $\hat{\alpha}$ is contained within the interval $0.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$, where α is the nominal Type I error rate (i.e., level of significance).

Analysis approaches

Before explaining the design of our study, we will briefly introduce three different approaches that are widely used for the analysis of repeated measures data,¹ and that we therefore included in our study. A detailed description of these procedures can be found in Keselman et al. (2001).

Univariate approach with df correction

In the one-factorial, completely within-subjects design that we simulated, each subject ($i = 1 \dots N$) is measured once under all levels ($k = 1 \dots K$) of the within-subjects factor, and there are no between-subjects (grouping) factors. In the *univariate approach*, the F -distributed test statistic is the ratio between the mean square for the within-subjects factor W and the mean square for the $W \times$ Subject interaction. This test statistic depends on the assumptions of normality, independence of subjects, and sphericity of the population variance–covariance matrix (Huynh & Feldt, 1970; Rouanet & Lépine, 1970). Sphericity means that all orthonormalized contrasts (e.g., $K - 1$ treatment differences) have the same population variance (Huynh & Feldt, 1970; Rouanet & Lépine, 1970). If the assumption of sphericity is violated, the F test will result in too many Type I errors. Box (1954) showed that in this case, the test statistic is approximately distributed as $F[\alpha; \varepsilon(K - 1); \varepsilon(N - 1)(K - 1)]$, where the population parameter ε quantifying the deviation from sphericity is in the interval $[1/(K - 1), 1.0]$ (Geisser & Greenhouse, 1958) and $\varepsilon = 1.0$ for a spherical population variance–covariance matrix (Huynh & Feldt, 1970).

Two popular sample estimates of ε have been proposed by Greenhouse and Geisser (1959) and Huynh and Feldt (1976). Both variants were included in our study, and will be denoted by GG and HF in the following. Other variants of df -adjusted univariate tests are discussed by Quintana and Maxwell (1994) and Hearne, Clark, and Hatch (1983).

Multivariate approach

The multivariate test of the effect of the within-subjects factor is performed by first creating $K - 1$ difference variables between pairs of factor levels, or more generally, $K - 1$ orthonormalized contrasts (for an excellent introduction, see Maxwell & Delaney, 2004). Next, Hotelling's (1931) multivariate T^2 statistic is used for testing the hypothesis that the vector of population means of these $K - 1$ difference variables equals the null vector. Unlike the univariate test, the

multivariate test (denoted by T^2 in the following) does not require sphericity, but only that the covariance matrix be positive definite. However, like the univariate approach, it is of course based on the assumptions of normality and the independence of observations across subjects. An important aspect is that the multivariate test requires $N \geq K$, which prevents its use if a small sample size is combined with a high number of factor levels, as was the case in most of the example studies cited above.

Mixed-model analysis

A third approach for testing the repeated measures main effect has gained importance during the last decade, partly due to the increasing availability of powerful computer hard- and software. This so-called *mixed-model analysis*² is often also referred to as *multilevel models* (especially in educational psychology and the social sciences), *hierarchical linear models*, or *random coefficient models* (especially if used with continuous covariates—that is, in multiple regression problems) (cf. Maxwell & Delaney, 2004, p. 763).

The mixed-model analysis is based on a linear model including both fixed and random effects. The random effects and the errors are assumed to be normally distributed. In contrast to the univariate and multivariate approaches, the variance–covariance structure of the response measure is modeled explicitly (Littell, Pendergast, & Natarajan, 2000); it depends on the variance–covariance structures of both the random effects and the errors. In fact, the univariate approach and the multivariate approach described above can be viewed as special cases of the mixed-model analysis. The PROC MIXED procedure from SAS 9.2 that we used for the simulations allows for fitting a wide variety of covariance structures (Wolfinger, 1996), among them are the spherical compound symmetry (CS) and the unstructured (UN) types (see Table 1).

Unlike the univariate and multivariate approaches, which use least-squares procedures, a maximum-likelihood or restricted maximum-likelihood approach (cf. Jennrich & Schluchter, 1986; Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006) is used for parameter estimation for the mixed model, with the potential consequence of numerical problems.

Note that the (restricted) maximum-likelihood-based PROC MIXED analysis can be used even if there are missing values (cf. Padilla & Algina, 2004)—more precisely, if the missing data mechanism is missing completely at random (MCAR) or missing at random (MAR) (Rubin, 1976). Missing values usually play no role for well-controlled

¹ Several other available procedures are targeted at analyzing the data from unbalanced designs (cf. Keselman et al., 2001; Vallejo & Livacic-Rojas, 2005; Vallejo Seco, Izquierdo, Garcia, & Diez, 2006). As we studied a completely within-subjects design without grouping factors, we did not include these procedures.

² Note that early texts referred to the univariate approach as a “mixed model,” because it contains a fixed effect of the within-subjects factor and a random effect of the subject (e.g., Geisser & Greenhouse, 1958).

Table 1 Population covariance structures used for data simulation

Compound symmetric (CS) (2 parameters)

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ & 1 & \rho & \rho \\ & & 1 & \rho \\ & & & 1 \end{bmatrix}$$

Unstructured (UN) ($K[K+1]/2$ parameters)

$$\begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1K} \\ & \sigma_{22}^2 & \sigma_{23} & \cdots & \sigma_{2K} \\ & & \sigma_{33}^2 & \cdots & \sigma_{3K} \\ & & & \ddots & \vdots \\ & & & & \sigma_{KK}^2 \end{bmatrix}$$

Random coefficients (RC) (4 parameters)

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & K-1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & K-1 \end{bmatrix} + \begin{bmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \ddots & \\ & & & \sigma^2 \end{bmatrix}$$

Heterogeneous first-order autoregressive [ARH(1)] ($K+1$ parameters)

$$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \cdots & \sigma_1\sigma_K\rho^{K-1} \\ & \sigma_2^2 & \sigma_2\sigma_3\rho & \cdots & \sigma_2\sigma_K\rho^{K-2} \\ & & \sigma_3^2 & \cdots & \sigma_3\sigma_K\rho^{K-3} \\ & & & \ddots & \vdots \\ & & & & \sigma_K^2 \end{bmatrix}$$

Only the upper diagonal of the symmetric matrices is shown. The CS structure is displayed for four factor levels ($K=4$). For higher K , the $K \times K$ correlation matrix again has 1 on the main diagonal and ρ elsewhere.

laboratory experiments, so we did not consider this issue in our study.

Previous simulation studies

Not many simulation studies measuring Type I error rates exist for completely within-subjects designs with nonnormal data and/or very small sample sizes. It has been suggested, however, that simulation results for designs with a combination of within- and between-subjects factors, identical group sizes (i.e., balanced designs), and equal covariance matrices across groups should be similar to what can be expected for a completely within-subjects design (e.g., Keselman et al., 2001, p. 11; Keselman, Kowalchuk, & Boik, 2000, p. 55). For “split-plot” designs containing both between- and within-subjects factors, many simulation studies are available (for reviews, see Keselman et al., 2001, 2002), due to the importance of this type of design for educational and clinical psychology and for the social sciences. Still, even if studies containing between-subjects factors are considered, it has to be stated that Type I error rates have not yet been obtained very systematically for extremely small sample sizes and nonnormal data. We will first discuss studies simulating small samples and nonnormal data, and then consider studies in which small sample sizes were simulated but the data were normally distributed.

For $N=6$ or 9 and a design with equal group sizes and equal covariance matrices across groups, Keselman et al. (2000) reported frequent liberal error rates if the data were $\chi^2(3)$ distributed. The multivariate approach showed a stronger tendency toward liberal error rates than did a univariate approach with df correction. For normally distributed data, both approaches controlled the Type I error rate (Huynh & Feldt, 1976; Lecoutre, 1991).

Berkovits, Hancock, and Nevitt (2000) simulated a design with a single within-subjects factor ($K=4$), no between-subjects factors, and varied the sample size between 10 and 60. For normal data and $N=10$, HF, GG, and T^2 controlled the Type I error rate. With increasing skewness and kurtosis of the simulated response variable, the three procedures produced conservative Type I error rates for a spherical population covariance matrix and showed liberal behavior at small values of ε . For strong deviations from normality (skewness = 3.0, kurtosis = 21.0), the results were robust only at the highest sample size ($N=60$).

Wilcox, Keselman, Muska, and Cribbie (2000) reported both conservative and liberal Type I error rates in a design with one within-subjects factor ($K=4$, $N=21$) if the distribution of the dependent variable was nonnormal. For strongly heavy-tailed distributions (g -and- h distribution with $h=0.5$; cf. Headrick, Kowalchuk, & Sheng, 2010), the univariate approach with Huynh–Feldt df correction often produced conservative Type I error rates, especially if the variances under the four factor levels were equal. If the variances were unequal, the multivariate approach produced conservative results for symmetric and heavy-tailed distributions, but highly liberal results for asymmetric and heavy-tailed distributions. In the latter case, HF also occasionally produced liberal Type I error rates.

Muller, Edwards, Simpson, and Taylor (2007) simulated normal data from a design without between-subjects factors ($K=9$), studied sample sizes between 10 and 40, and varied the sphericity parameter ε of the population covariance matrix. At an α level of .04, GG and HF always met Bradley’s liberal criterion. PROC MIXED fitting a UN or CS type of covariance matrix and using the Kenward–Roger (1997) adjustment produced liberal Type I error rates if $N=10$ and $\varepsilon < 1.0$. At $N=20$, PROC MIXED produced robust results if a UN rather than a CS type of covariance matrix was fitted.

Skene and Kenward (2010a) simulated a design with two groups, five subjects per group (i.e., $N = 10$), and a continuous within-subjects covariate with five levels ($K = 5$). The response variable was normally distributed. PROC MIXED fitting an unstructured covariance matrix and using the Kenward and Roger (1997) adjustment controlled the Type I error rate.

Gomez, Schaalje, and Fellingham (2005) simulated normal data, with one continuous within-subjects factor (three to five levels), one between-subjects factor with three levels, and total sample sizes between nine and 15. For a balanced design with identical covariance matrices across groups, PROC MIXED with the Kenward–Roger adjustment controlled the Type I error rate if the correct covariance structure was fitted, except for a single type of population covariance structure. If the covariance structure was selected via the Akaike (AIC; Akaike, 1974) or Bayesian (BIC; Schwarz, 1978) information criteria, several liberal error rates were reported.

Taken together, for samples sizes as low as $N = 6$, the univariate approaches with *df* correction and the multivariate approach seem to control Type I error rates for normal data, which is the expected result. With some exceptions, this is also true for PROC MIXED with the Kenward–Roger adjustment. However, this pattern seems to change, sometimes even dramatically, if the distribution of the response variable is nonnormal.

To summarize the previous results, for repeated measures designs, no empirical Type I error rates have been reported for $N < 6$, the influence of sample size has only been investigated in most studies by comparing two or three different sample sizes, and the effects of nonnormality have also not been tested for very small sample sizes combined with higher numbers of levels of the within-subjects factor. Thus, our study for the first time provides empirical Type I error rates for (1) samples sizes as small as three and varied in small steps, (2) normal and nonnormal data, (3) numbers of factor levels ranging from four to 16, and (4) a rather wide variety of population covariance matrices.

Method

Simulated experimental design

We simulated a completely within-subjects design, in which each subject ($i = 1 \dots N$) was measured once under all levels ($k = 1 \dots K$) of the within-subjects factor (i.e., each experimental condition). There were no between-subjects (grouping) factors. An example of such a design would be a Stroop color-word task (cf. MacLeod, 1991) in which, for each subject, response times are measured for incongruent, congruent, and neutral trials (i.e., $K = 3$). As we were interested in empirical Type I error rates, we simulated data sets corresponding to the

null hypothesis of equal means—that is,

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k, \quad (1)$$

where μ_k is the population mean in condition k .

Characteristics of the simulated data sets

The simulated data sets varied in (a) the population distribution, (b) the population covariance structure, (c) the number of levels of the within-subjects factor, and (d) the sample size (i.e., the number of simulated subjects). These variables were factorially combined, and each resulting scenario was simulated 5,000 times. It should be noted that PROC MIXED requires substantial computation time, especially at large values of K and N when many parameters have to be estimated.

Population distributions The following distributional forms of the simulated response variable were chosen for this study: (1) a normal distribution with $\mu = 0$ and $\sigma = 1$, (2) a log-normal distribution³ with $\mu = 0$ and $\sigma = 1$, and (3) a chi-squared distribution with two degrees of freedom.

The selected log-normal distribution had skewness and kurtosis values of 6.18 and 113.94, respectively. Skewness is a measure of asymmetry, and a distribution with positive kurtosis has a heavier tail and a higher peak than the normal distribution (see DeCarlo, 1997). Thus, the simulated log-normal distribution exhibited a degree of asymmetry and heavy-tailedness that, according to previous studies, might negatively affect the performance of the analysis procedures (Sawilowsky & Blair, 1992; Wilcox et al., 2000). Most previous simulation studies had used a log-normal distribution with $\mu = 0$ and $\sigma = 0.5$ (e.g., Algina & Oshima, 1994), which has smaller values of skewness and kurtosis (1.75 and 8.89, respectively).⁴ We selected a distribution showing a rather extreme deviation from normality in order to test the robustness of the data analysis approaches in such cases. Note that skewed distributions are frequently encountered in psychological studies (Micceri, 1989; Wilcox, 2005), although it is unclear which values of, for instance, skewness are representative for a particular field of research (Keselman, Kowalchuk, Algina, Lix, & Wilcox, 2000; Wilcox, 2005). As we discussed above, the log-normal distribution can be considered a first approximation to the distribution of response times (cf. Heathcote et al., 2004), which is a dependent measure very frequently used in cognitive psychology. Recently,

³ The probability density function for the log-normal distribution is $P(x) = \frac{1}{\sigma\sqrt{2\pi x}} e^{-(\ln x - \mu)^2 / (2\sigma^2)}$

⁴ Several articles have erroneously claimed that a log-normal distribution with $\mu = 0$ and $\sigma = 0.5$ has a kurtosis of 5.90 rather than 8.89. This error can be traced back to Keselman, Algina, Kowalchuk, and Wolfinger (1999b, p. 71).

Palmer, Horowitz, Torralba, and Wolfe (2011) reported empirical values of skewness and kurtosis for response time distributions from a visual search experiment. They found skewness values greater than 4.0 and kurtosis values greater than 40.0 in some conditions, which again justifies our decision to include a log-normal distribution with rather high values of skewness and kurtosis.

The $\chi^2(2)$ distribution exhibits extreme asymmetry (Micceri, 1989), because its mode coincides with the minimum. Although the values of skewness = 2 and kurtosis = 6 are moderate, the $\chi^2(2)$ distribution is not only asymmetric but has a monotonically decreasing probability density function (PDF). Note that a $\chi^2(2)$ distribution multiplied by a constant, as is included in the covariance matrices with unequal variances, is strictly speaking no longer a $\chi^2(2)$ distribution, although in the literature it is usual to ignore this issue and to refer only to the base distribution used when generating the random variates (e.g., Keselman, Carriere, & Lix, 1993). Incidentally, a binomial distribution with success rate $p = .05$ —as would, for example, occur for error rates in many experiments in cognitive psychology—also shows a strictly decreasing PDF if each proportion is based on fewer than 20 trials, as in, for example, the study by Meiran (1996) cited above.

Population covariance structures As we discussed in the introduction, an important concern for repeated measures ANOVAs is the sphericity (or, rather, lack of sphericity) of the population covariance matrix (e.g., Huynh & Feldt, 1970; Rouanet & Lépine, 1970). For example, the univariate ANOVA approach assumes sphericity (Huynh & Feldt, 1976), and the correction factors for the degrees of freedom (e.g., Greenhouse & Geisser, 1959; Huynh & Feldt, 1976) attempt to correct for deviations from a spherical pattern, as quantified via Box's ε (Box, 1954), which is a population parameter. For this reason, we included a compound symmetric (CS) population covariance structure (see Table 1), which is spherical (i.e., $\varepsilon = 1.0$). In addition, three population covariance structures with $\varepsilon = .5$ were studied, thus showing a strong deviation from sphericity. As is shown in Table 1, these structures were of the unstructured (UN), random-coefficients (RC), and heterogeneous first-order autoregressive [ARH(1)] types (cf. Wolfinger, 1996). The reason to include different types of covariance structures with $\varepsilon = .5$ was to evaluate the performance of PROC MIXED with different fitted covariance structures. Beyond that, we were interested in whether the Type I error rates of the other ANOVA procedures would depend on the specific covariance structure for equal values of ε .

For $K = 4$ and $\varepsilon = .75$, covariance matrices of types UN, RC, and ARH(1) were specified by Keselman, Algina, Kowalchuk, and Wolfinger (1999a). These matrices were subsequently used in a substantial number of simulation studies. To construct covariance matrices for a higher

number of factor levels and with $\varepsilon = .5$, we created covariance matrices with the appropriate structures, using randomly selected values for the parameters. For each pairing of covariance structure and number of factor levels, we then selected a covariance matrix with ε close to the desired value of .50 (squared deviation $\leq 10^{-4}$). For the power-transform method (Headrick, 2002) that we used to generate multivariate nonnormally distributed data (see the Data Generation section), positive definite covariance matrices are required because the method involves a Cholesky decomposition (Commandant Benoit, 1924). Additionally, the intermediate covariance matrices used by the power-transform method (see the Data Generation section) are also required to be positive definite. We selected covariance matrices to meet these criteria.

As can be seen in Table 1, the RC covariance structure needs four parameters: σ , σ_{11} , σ_{22} , and σ_{12} . This is true for all numbers of factor levels. The common standard deviation σ was randomly selected from an interval from 0.1 to 4. The parameters σ_{11} , σ_{22} , and σ_{12} were sampled uniformly from an interval from -0.9 to 0.9 .

The ARH(1) covariance structure has only one correlation parameter (ρ) and a variance parameter for every factor level (σ_i^2). Therefore, the ARH(1) covariance structure needs $K + 1$ parameters. The standard deviations σ_i were first uniformly sampled from an interval from 1.0 to 4.0, and then the correlation parameter ρ was selected to produce the desired $\varepsilon = .5$.

The UN covariance structure has $K(K + 1)/2$ parameters. While K of these parameters represent the variances of the K variables, the remaining parameters represent the covariances. The standard deviations σ_{ii} were first sampled uniformly from an interval from 0.5 to 6. Then the covariances σ_{ij} were sampled uniformly from the interval $[0.30\sigma_i\sigma_j, 0.97\sigma_i\sigma_j]$.

The CS covariance structure needs only two parameters, independent of the number of factor levels. This covariance structure is distinct from the other three, because by design the Box epsilon is $\varepsilon = 1.0$ for all CS covariance matrices (i.e., the matrix is spherical) and because all variances are equal. We used the parameter values $\sigma^2 = 20.0$ and $\rho = .8$.

Supplement B (available online with this article) shows the covariance matrices used for data generation. Note that the nonspherical matrices also exhibited heterogeneous variances for the K variables.

Numbers of levels of the within-subjects factor For the number of levels of the within-subjects factor (K), the following values were chosen: 4, 8, and 16. While a value of $K = 16$ might seem high, experiments using, for example, a large number of presentation levels (e.g., Florentine, Buus, & Poulsen, 1996) are not unusual in psychophysics. Additionally, a test with $16 - 1 = 15$ numerator degrees of freedom would arise when testing for the interaction between two within-subjects factors that have four and six

factor levels, respectively. The number of factor levels has been demonstrated to affect the Type I error rate and the power of repeated measures ANOVAs (e.g., Algina & Keselman, 1997; Keselman, Keselman, & Lix, 1995). Additionally, some of the analysis procedures cannot be applied if a high number of factor levels is combined with a small sample size. For example, the multivariate approach requires that the number of subjects N be equal to or greater than the number of factor levels of the within-subjects factor ($N \geq K$).

Sample sizes As we were especially interested in the performance of the procedures for extremely small sample sizes, we varied the sample size in steps of one between $N = 3$ and 10, and we also studied nine larger sample sizes ($N = 16, 17, 18, 19, 20, 30, 50, 70,$ and 100). The small steps between 16 and 20 were included because pilot simulations indicated that PROC MIXED needs slightly more subjects than the number of factor levels for convergence when fitting a UN covariance structure.

Data analysis approaches

The empirical Type I error rates were analyzed for four different repeated measures ANOVA procedures. These were (1) the univariate approach with Greenhouse–Geisser correction for the degrees of freedom (GG), (2) the univariate approach with Huynh–Feldt df correction (HF), (3) the multivariate approach (T^2), and (4) the mixed-model approach as computed by SAS PROC MIXED (see Keselman et al., 2001, for a description of the different approaches). Note that for the univariate approaches, the df correction was unconditionally applied (Keselman, Rogan, Mendoza, & Breen, 1980), instead of first conducting a test for lack of sphericity (Mauchly, 1940). The primary reason for this was that the Mauchly test requires $N \geq K(K - 1)/2$, which rendered it useless for a considerable proportion of our simulated designs. Beyond that, the Mauchly test is sensitive to departures from normality (Huynh & Mandeville, 1979).

The analyses in PROC MIXED were conducted using the Kenward–Roger (1997)⁵ adjustment (option `ddfm=KR` in SAS 9.2), which has been demonstrated to be superior to alternative methods of computing the degrees of freedom (Arnau, Bono, & Vallejo, 2009; Fouladi & Shieh, 2004; Kowalchuk, Keselman, Algina, & Wolfinger, 2004; Schaalje, McBride, & Fellingham, 2002; Skene & Kenward, 2010a). Additionally, for the SAS PROC MIXED analyses, several different model covariance structures were fitted (cf. Keselman et al., 1999b; Kowalchuk et al., 2004; Littell et al., 2000)—

namely UN, CS, ARH(1), CSH, HF, and RC (see Table 1 and Wolfinger, 1996, for detailed specifications of these structures). It should be noted that PROC MIXED offers a large amount of flexibility for specifying covariance structures. For example, the CS structure can be fit in two different ways (cf. Littell et al., 2000). The SAS syntax that we used for the six variants is available as supplemental material (Supplement A) in the journal's electronic supplementary archive.

Data generation

We simulated multivariate normally distributed data using the method of Kaiser and Dickman (1962), and nonnormal correlated data via the power method transformation proposed by Headrick and coworkers (Headrick, 2002; Headrick & Kowalchuk, 2007; Headrick, Sheng, & Hodis, 2007). The power method is based on work by Fleishman (1978) and by Vale and Maurelli (1983). Fleishman proposed a moment-matching approach to simulate nonnormally distributed data. First, normal deviates are generated. Then a polynomial transform of order three is applied, with the polynomial coefficients selected so that the first four moments of the transformed data match the corresponding moments of the target distribution. Headrick (2002) enhanced this approach by using fifth-order polynomials. Therefore, it is possible to match the first six moments of a distribution. It is also possible to match a wider range of distributions (e.g., combinations of skewness and kurtosis).

Following Headrick (2002) and Headrick et al. (2007), our simulations started by generating K independent standard normal deviates ($X = X_1 \dots X_K; \mu = 0, \sigma = 1$) with the SAS IML function `RANDNORMAL`. From these independent normal deviates, K correlated normal deviates $Z = Z_1 \dots Z_K$ with population correlation matrix R^* were generated using the method of Kaiser and Dickman (1962). For the $K \times K$ matrix R^* , the Cholesky decomposition $U'U = R^*$ (where U is an upper triangular matrix and U' is the transpose of U) yielded $K(K + 1)/2$ coefficients. These coefficients were used to compute K linear combinations of each X_k , $Z = UX$, which resulted in Z , correlated with the population correlation matrix R^* (for an example, see Headrick et al., 2007, p. 11; Kaiser & Dickman, 1962). As Vale and Maurelli (1983) have shown, if a polynomial transformation is applied to these correlated normal deviates, the correlations between the resulting non-normal deviates will not be equal to the correlations between the normal deviates that they were based on. Therefore, to produce K correlated nonnormal deviates with the intended population correlation matrix R , the Z_k are generated with an intermediate correlation matrix R^* that was selected so that, after applying the polynomial transformations to the Z_k , the target correlation matrix R results. Headrick (2002) noted that if the correlation between a pair of standard normal deviates

⁵ Kenward and Roger (2009) recently suggested an improved variant of their procedure. However, because we are not aware of a statistical package incorporating the improved algorithm, the results reported in this article apply to the original procedure by Kenward and Roger (1997), which was used in SAS Version 9.2.

$(Z_i$ and $Z_j)$ is $\rho_{Z_i Z_j}$ and then a polynomial transformation is applied to each of the two correlated standard normal deviates, so that two power-transformed variables (Y_i and Y_j) with mean 0 and variance 1 result, then the correlation $\rho_{Y_i Y_j}$ is given by the expected value of the product of the latter two random variables, $\rho_{Y_i Y_j} = E(Y_i \cdot Y_j)$ (see Eq. 26 in Headrick, 2002). To determine the intermediate correlation matrix \mathbf{R}^* , this equation is solved for $\rho_{Z_i Z_j}$, separately for each of the pairwise correlations, and given the to-be-applied polynomial coefficients (see Eq. 2 below) and the target correlations $\rho_{Y_i Y_j}$. We solved the equation numerically with the Wolfram Research Mathematica function *NMinimize*. The values of $\rho_{Z_i Z_j}$ constitute the intermediate correlation matrix \mathbf{R}^* .

In the next step, the correlated normal deviates Z_k were transformed into nonnormal deviates Y_k , each with mean 0 and variance 1, using the polynomial transformation

$$Y_i = c_{1i}Z_i + c_{2i}Z_i^2 + c_{3i}Z_i^3 + c_{4i}Z_i^4 + c_{5i}Z_i^5 + c_{6i}Z_i^6, \quad i = 1 \dots K. \quad (2)$$

The polynomial coefficients c_{1i} through c_{6i} were chosen so that the first six standardized cumulative moments of Y_k matched the first six moments of the target distribution— $\chi(2)$ or LogNormal[0, 1] (Headrick, 2002). The moments of a power-transformed normal deviate are available in closed form (cf. Headrick, 2002). We used a numerical procedure (Wolfram Research Mathematica *NMinimize*) to find the coefficients minimizing the sum of squared deviations between the six moments of the power-transformed normal deviate and the corresponding six moments of the target distribution [e.g., $\chi(2)$], with an accuracy goal of 10^{-20} .

The Y_k are nonnormal deviates with mean 0 and variance 1, correlated with the desired population correlation matrix \mathbf{R} . To create correlated nonnormal deviates with the desired population covariance matrix $\mathbf{\Sigma}$, in the final step, each Y_k was multiplied by the desired standard deviation. Note that we simulated data corresponding to the null hypothesis $\mu_1 = \mu_2 = \dots = \mu_K = 0$.

To check the quality of our simulations, we computed the empirical covariance matrix, as well as the empirical moments of the individual random variables, for each of the 36 simulated multivariate distributions (i.e., for each combination of distribution of the simulated response variable, covariance structure, and number of factor levels). For each of these conditions, 1,959,999 data sets were simulated. The maximal deviation of the empirical Box (1954) ε from the target ε (1.0 or .5) was 0.51 %. The average root-mean squared deviation of the variances and covariances in the upper diagonal of the empirical covariance matrix, as compared to the intended covariance matrix (see Supplement B), was 0.0047, with a maximum of 0.101.

The first four empirical standardized moments (mean, variance, skewness, and kurtosis) were also found to be within a

narrow range around the target values. Only for the log-normal distribution did the kurtosis show rather large variability, which is to be expected, because the fourth sample moment about the mean is strongly affected by extreme values.

Taken together, the empirical covariance matrices and the empirical moments showed that the simulation algorithm worked as expected.

Simulation program

The simulation program was written in the SAS MACRO and SAS/IML languages and was run on version 9.2 of SAS.

Results

Type I error rates

For each simulated data set, the p value produced by a given analysis approach was compared to the nominal α level of .05, and the empirical Type I error rate $\hat{\alpha}$ for a condition (e.g., $N = 3$, $K = 4$, normally distributed data, CS population covariance structure) was computed as the number of tests in which $p < \alpha$, divided by the total number of p values for this condition (i.e., 5,000, if all tests produced a p value). The empirical Type I error rates are binomially distributed; thus, the 95 % Wald confidence interval for the proportion $\hat{\alpha}$ is $\hat{\alpha} \pm 1.96 \times \sqrt{\hat{\alpha}(1 - \hat{\alpha})/N}$ where N is the number of simulations. If, for example, the observed Type I error rate was .05 (i.e., 250 tests out of 5,000 were significant at the nominal α level of .05), then the true Type I error rate could be expected to be in the interval [.0456, .0544] with a probability of 95 %.

For the PROC MIXED analyses, we analyzed the empirical Type I error rates in five different cases: (1) fitting the correct covariance structure—for instance, fitting an UN structure if the data were generated using a population covariance matrix of type UN; (2) selecting the best PROC MIXED model in terms of the AIC (Akaike, 1974); (3) selecting the model on the basis of the BIC (Schwarz, 1978); (4) unconditionally fitting an unstructured (UN) model covariance matrix; or (5) unconditionally fitting a covariance matrix of the “heterogeneous compound symmetry” (CSH) type. In the following discussion, PROC MIXED Variants 1–5 are denoted by PM_{CC} , PM_{AIC} , PM_{BIC} , PM_{UN} , and PM_{CSH} , respectively. Thus, as, for example, in Kowalchuk et al. (2004), one variant of the PROC MIXED approach used prior knowledge about the population covariance structure. It is, of course, unlikely that a researcher will know in advance which covariance structure his or her data will exhibit, but fitting the correct covariance structure should represent the best possible performance of the mixed-model approach. The two variants selecting the best covariance structure using likelihood-based information

criteria attempted a trade-off between global goodness of fit and the number of model parameters (cf. Keselman et al., 2002). Thus, at least in theory, these information criteria should select the most parsimonious model still showing an acceptable level of goodness of fit. Finally, one of the two remaining variants always assumed a UN covariance structure, which should incorporate all possible empirical covariance structures, because it places no constraints on the structure of the covariance matrix. This flexibility comes at the cost of having to estimate a large number of parameters, making it impossible to use if a high number of factor levels is combined with a small sample size, and also potentially causing problems with numerical convergence (cf. Gomez et al., 2005). The CSH structure, on the other hand, is in a sense located between the UN structure and the CS structure (for a discussion, see, e.g., Wolfinger, 1996). The CS structure has equal variances and equal correlations. The CSH structure allows for unequal variances (i.e., entries on the main diagonal), while still assuming a constant correlation between all pairs of factor levels (Wolfinger, 1996). Note that the CSH structure is somewhat similar to the HF structure (Wolfinger, 1996), the latter corresponding to the spherical “Type H” structure discussed by Huynh and Feldt (1970). We opted for the CSH rather than the HF structure because, in our experience with analyzing real data sets, PROC MIXED models using the HF covariance structure frequently show problems with convergence, while fitting the CSH structure seems less problematic. This assumption was clearly supported by the convergence rates observed in our simulations (see the Convergence Rates section below).

Normally distributed data The empirical Type I error rates for normally distributed data are displayed in Fig. 1. Data points above the red or below the blue horizontal line indicate liberal or conservative error rates, respectively, according to Bradley’s liberal criterion. A procedure can be considered robust if all data points are located between the blue and red horizontal lines. Note that in Fig. 1 the filled circles are for nonspherical population covariance matrices with homogeneous variances (denoted by HV); these data will be discussed in the following section.

Across the 204 combinations of population covariance structure [ARH(1), CS, RC, and UN], number of factor levels, and sample size, the univariate approach with Huynh–Feldt *df* correction (HF) controlled the Type I error rate according to Bradley’s (1978) liberal criterion, except for four cases characterized by $K = 16$ and $N < 7$, for which the error rate was slightly higher than .075. At the smaller samples sizes, the error rates had a small tendency toward liberal error rates for the nonspherical population covariance structures, and a smaller tendency toward conservative error rates for the spherical CS structure. As expected, the GG *df* correction frequently resulted in conservative error rates at

the smaller sample sizes, especially for the spherical CS population covariance matrix (open circles in Fig. 1; cf. Huynh & Feldt, 1976). The multivariate approach (T^2) always showed almost perfect control of the error rate, which was expected because its assumptions were all met. However, T^2 is applicable only for $N \geq K$. When fitting a UN covariance structure, PROC MIXED required $N > K$ for convergence (see the Convergence Rates section below) and produced liberal error rates for $N = K + 1$. The latter result was also obtained by Muller et al. (2007). Note that except for these conditions, the error rates were identical to the error rates produced by the multivariate approach, which is the expected result (Skene & Kenward, 2010a). When fitting the CSH structure, PROC MIXED frequently produced conservative error rates at small sample sizes and some liberal error rates at higher sample sizes. Selecting the covariance structure via AIC or BIC worked reasonably well for normally distributed data. At $K = 4$, both variants produced some liberal error rates at $N < 10$. Additionally, selection via AIC or BIC produced liberal error rates if $N = K + 1$, which can be attributed to the liberal behavior of PROC MIXED fitting a UN structure in this condition. This problem did not occur for selection via BIC if $K = 16$. Some error rates were conservative at $N = 3$ or 4, especially for selection via BIC. Finally, if the correct covariance structure was fit, the error rates were close to the nominal values, except for the two smallest samples sizes when the population covariance structure was ARH(1), and of course liberal rates for the UN population covariance structure at $N = K + 1$.

Taken together, for normally distributed data the HF test controls the Type I error rate with only very few exceptions and can therefore be recommended, especially if extremely small sample sizes are combined with high numbers of factor levels of the within-subjects factor. The multivariate approach worked perfectly if $N \geq K$. For these sample sizes, future studies should compare the power of the HF and the T^2 tests. Results by Algina and Keselman (1997) have suggested that the multivariate approach would be more powerful than HF if the sample size was large, the deviation from sphericity was large (i.e., $\epsilon < .85$), and the number of factor levels was small. PM_{AIC} , PM_{BIC} , and the benchmark variant PM_{CC} performed reasonably well, but all of them occasionally produced conservative or liberal Type I error rates at sample sizes smaller than 20. For $N > K + 1$, the error rates for PM_{UN} were identical to those produced by the multivariate approach. PM_{CSH} produced many nonrobust results and can therefore not be recommended.

$\chi^2(2)$ -distributed data Figure 2 shows the Type I error rates for $\chi^2(2)$ -distributed data. The results indicate a strong effect of the combination of population covariance matrix and number of factor levels, compatible with previous findings (Berkovits et al., 2000). For the spherical CS matrix,

there was a general tendency toward conservative error rates, increasing with the number of factor levels. This problem was most pronounced for PROC MIXED fitting a CSH structure, followed by the GG and the HF tests. The multivariate approach also generally produced slightly conservative error rates for the CS covariance structure, but in most cases met Bradley's liberal criterion. PROC MIXED with covariance structure selection via AIC or BIC controlled the error rate for $N > K + 1$. PROC MIXED produced robust results when fitting the correct covariance structure.

In contrast to this generally conservative behavior observed for the CS population covariance matrices, for the nonspherical population covariance matrices the error rates were often liberal if the data were $\chi^2(2)$ distributed. In fact, at $K = 4$ all procedures produced liberal error rates, except for the univariate *df*-adjusted approaches at the highest sample sizes. At $K = 8$, the univariate *df*-adjusted procedures controlled the Type I error rate except at the smallest sample sizes, where the GG test again erred on the conservative side, and HF produced some liberal error rates. The multivariate approach and PROC MIXED produced liberal results. At $K = 16$, the HF test showed robust behavior for the nonspherical population covariance matrices, while the GG procedure often behaved conservatively, and the error rates produced by the remaining procedures were frequently liberal. It should be noted that even at the three highest sample sizes ($N \geq 50$), only GG and HF behaved robustly, while the problems with nonrobust Type I error rates remained for the other procedures. Across all conditions, the error rates were closest to the nominal value for procedure HF, but even this procedure produced many liberal or conservative error rates according to Bradley's liberal criterion.

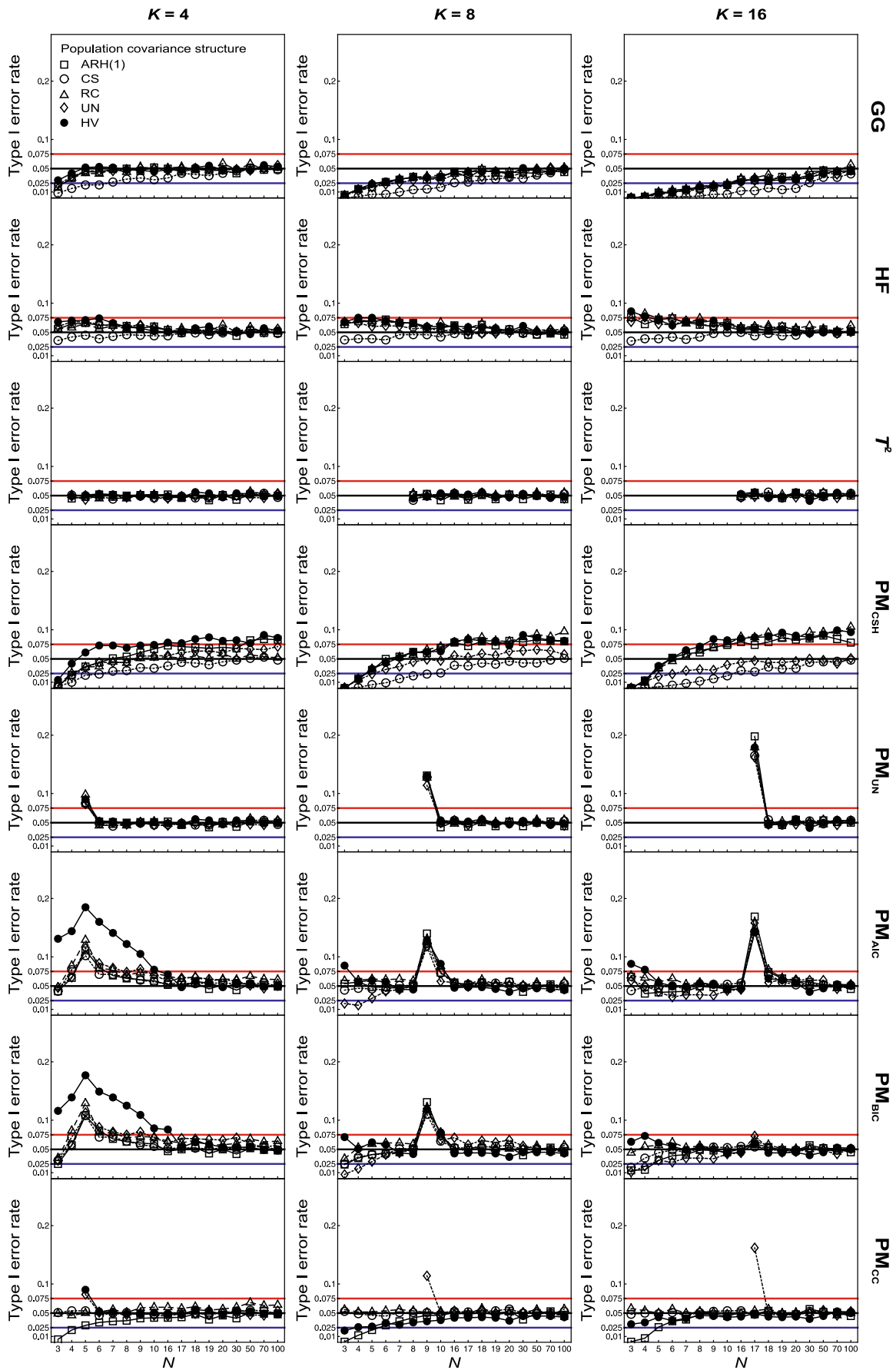
Log-normally distributed data Figure 3 shows the Type I error rates for log-normally distributed data. In contrast to the results for the $\chi^2(2)$ -distributed data, here there was a general tendency toward conservative error rates. At the two largest sample sizes ($N = 70$ or 100), several liberal values were observed. The number of conservative values was higher for the CS than for the nonspherical population covariance structure and increased with the number of factor levels. The error rates produced by the T^2 test were more frequently within Bradley's (1978) interval than for the *df*-corrected univariate approaches. Surprisingly, the number of nonrobust results was nonmonotonically related to the sample size for the multivariate approach, with two peaks at intermediate and large sample sizes (see Fig. 6 below). At the smaller sample sizes where the multivariate test was not applicable, PROC MIXED with covariance structure selection via AIC or BIC showed slightly better control of the Type I error rate than did the univariate approaches, although many values were still conservative according to Bradley's liberal criterion. Additionally, because PM_{UN}

Fig. 1 Normal data: Empirical Type I error rates for the eight analysis procedures (rows), as a function the number of factor levels (columns), sample size (N), and the population variance–covariance structure. Data points located between the red and blue lines are considered robust according to Bradley's (1978) liberal criterion. The symbols indicate the different population covariance structures: boxes, ARH(1); open circles, CS; triangles, RC; diamonds, UN; and filled circles, HV (nonspherical population covariance structure with homogeneous variances; see the following section). GG and HF indicate univariate approaches with Greenhouse–Geisser and Huynh–Feldt *df* corrections, respectively; T^2 , the multivariate approach; PM_{CSH} and PM_{UN} , PROC MIXED fitting a CSH or a UN covariance structure, respectively; PM_{AIC} and PM_{BIC} , PROC MIXED with model covariance structures selected via information criteria; and PM_{CC} , PROC MIXED fitting the correct (population) covariance structure

again produced liberal results for $N = K + 1$, PM_{AIC} and PM_{BIC} also showed liberal error rates in most of these conditions. Notably, the empirical error rates for PROC MIXED with covariance structure selection via information criteria were often closer to the nominal value than if the correct covariance structure was fitted, especially at $K = 4$.

Nonspherical population covariance structures with homogeneous variances

The results showed a pronounced effect of the population covariance structure for $\chi^2(2)$ -distributed data. Can the observed differences between the error rates for the CS structure and for the remaining structures be attributed to the deviation from sphericity (i.e., $\varepsilon = 1.0$ vs. $\varepsilon = .5$), or to the fact that the K variances were equal for the CS population covariance matrix but unequal for the other matrices? To gain a preliminary insight into this question, we decided to conduct additional simulations for population covariance structures with homogeneous variances but $\varepsilon = .5$. We denote these structures by HV, for “homogeneous variances.” For $K = 8$ or 16 , we used a first-order autoregressive correlation structure. The correlations in this structure follow the same pattern as for the ARH(1) matrix depicted in Table 1, but the variances were all identical and equal to 16.0. For $K = 8$ and $K = 16$, the matrix has $\varepsilon = .5$ if $\rho = .760$ or $.632$, respectively. For $K = 4$, it is not possible to construct an AR(1) matrix with $\varepsilon = .5$. Therefore, we created a matrix with all variances set to 16.0 and with the correlations randomly sampled from the intervals $[-.9, -.3]$ and $[.3, .9]$. We denote this matrix by UN_s , for “unstructured–same variance.” The population covariance matrices are displayed in Supplement B. The simulation results for the nonspherical matrices with equal variances are shown by the filled circles in Figs. 1–3. For normal data (Fig. 1), the Type I error rates did not differ strongly between the nonspherical matrices with homogeneous versus heterogeneous population variances. At $K = 4$, PM_{AIC} , PM_{BIC} , and PM_{CSH} produced more liberal error rates at small sample sizes with the HV population covariance structure than with the remaining



structures. At $K = 8$, procedure HF produced a liberal result at $N = 4$ or 5 for the AR(1) population covariance structure, but not for the nonspherical structures with heterogeneous population variances. PM_{CSH} was liberal at larger sample sizes for the AR(1) but not for the UN population covariance structure, and PM_{AIC} and PM_{BIC} both produced one additional liberal error rate with the AR(1) structure. At $K = 16$, the AR(1) structure resulted in one additional liberal error rate for HF (at $N = 3$), and in liberal error rates for PM_{CSH} at large sample sizes.

For the $\chi^2(2)$ -distributed response variable (Fig. 2), at $K = 4$ GG, HF, T^2 , PM_{UN} , and PM_{CC} showed robust behavior for the UN_s population covariance structure, with only four exceptions, in contrast to the generally liberal error rates observed with the population covariance matrices with heterogeneous variances. PM_{AIC} and PM_{BIC} also controlled the Type I error rate with the UN_s structure if $N > 10$. At $K = 8$, where the error rates of most procedures were liberal with heterogeneous population variances, the AR(1) structure removed this tendency toward liberal decisions, resulting in robust behavior for HF, T^2 , PM_{UN} , PM_{AIC} , and PM_{BIC} in most conditions. Essentially the same pattern was observed at $K = 16$. Thus, for the $\chi^2(2)$ -distributed response variable, nonspherical population covariance matrices with homogeneous variances removed the strong tendency toward liberal results observed for the population covariance structures with heterogeneous variances. However, the error rates for the nonspherical population covariance matrices with homogeneous variances were not identical to the results for the CS structure, which shows variance homogeneity and covariance homogeneity, and for which the error rates were often conservative.

For the log-normal data (Fig. 3), the Type I error rates generally tended to be lower for the nonspherical population covariance matrices with homogeneous variances, as compared to the conditions with heterogeneous population variances. This resulted in more frequent conservative error rates at smaller sample sizes, but also in the absence of liberal error rates at the three largest sample sizes. Again, the error rates for the nonspherical population covariance matrices with homogeneous variances were not identical to the results for the CS structure.

Taken together, these results indicate that both variance heterogeneity and correlation heterogeneity affect the Type I error rates for nonnormal data, and this finding is compatible with the results by Wilcox et al. (2000). Heterogeneous variances probably present a stronger problem than do heterogeneous covariances. Additional studies will be desirable here. On a more general level, our study shows that it is worthwhile to test a wider variety of population covariance structures than those included in previous studies.

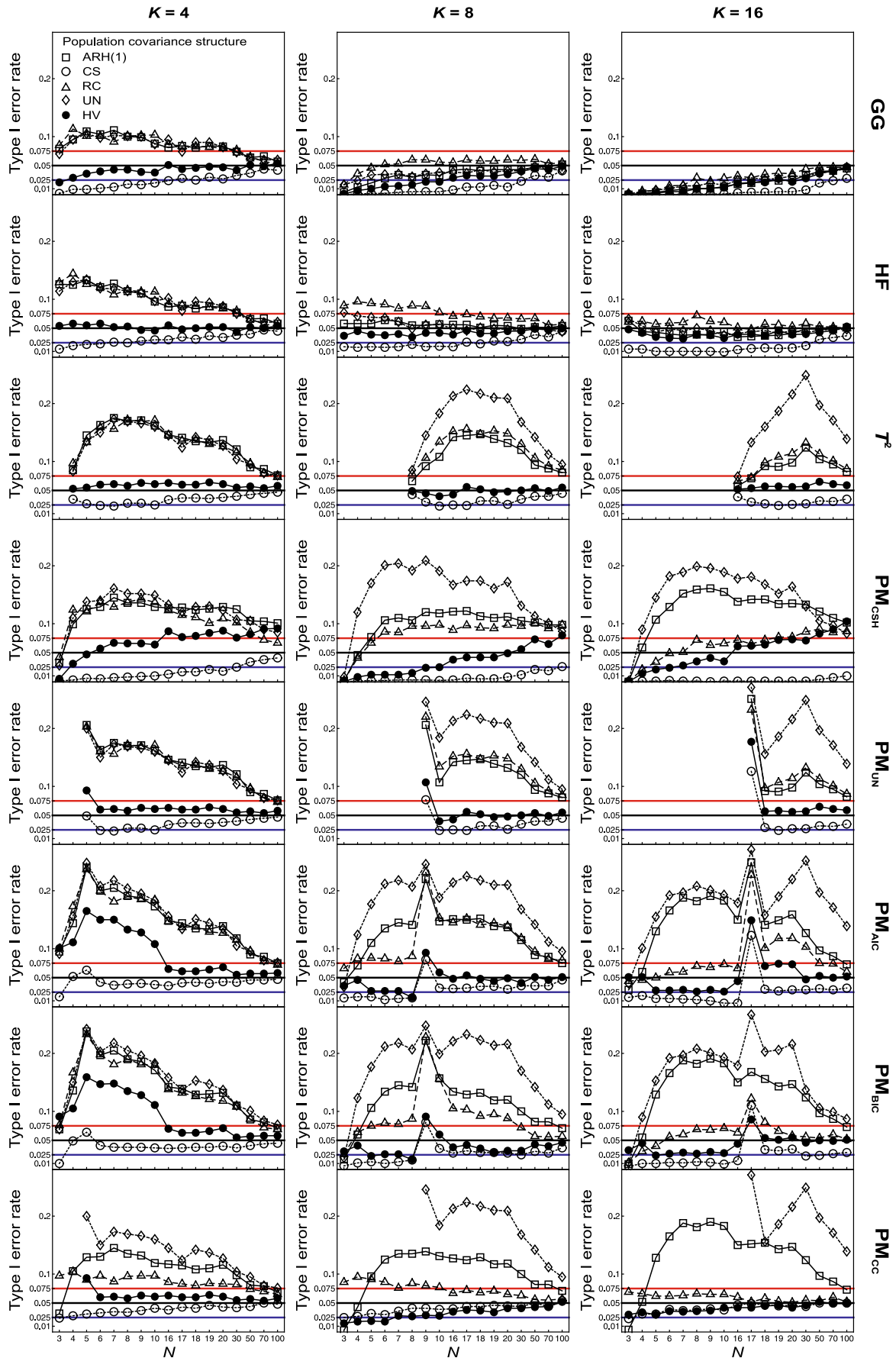
Fig. 2 Chi-square-distributed data: Empirical Type I error rates for the eight analysis procedures (rows), as a function the number of factor levels (columns), sample size (N), and the population variance–covariance structure (symbols). The format is the same as in Fig. 1

Convergence rates

The univariate and multivariate approaches are based on least-squares estimation and always produce a p value for the test of the omnibus hypothesis, given that the sample size is high enough to provide sufficient residual dfs . In contrast, the mixed-model analyses based on (restricted) maximum-likelihood can exhibit problems with convergence, and thus fail to produce a p value for a given data set. While most previous studies have ignored this problem (but see Gomez et al., 2005), for researchers planning to apply this specific type of analysis, it is important to know how likely these convergence problems will be for a given condition. For this reason, we report the proportions of tests in which a particular analysis procedure failed to produce a p value.

As can be seen in Supplement C, PROC MIXED always converged when fitting the CS model covariance structure. For the ARH type, the convergence rate was higher than .98 in all conditions. With the RC type, we observed some convergence rates smaller than .95 at $N = 3$, but with only a single exception, the convergence rates were higher than .90 even at this smallest sample size. For the CSH type, convergence rates higher than .95 were obtained except at the smallest samples sizes ($N \leq 4$ for $K = 4$, $N \leq 5$ for $K = 8$, and $N \leq 6$ for $K = 16$). For the UN model covariance structure, the data indicated that the convergence rate was close to 1.0 as soon as the sample size was larger than the number of factor levels ($N > K$). Compatible with our own experience when analyzing real data, the problems with convergence were most pronounced when fitting the HF type model covariance structure. Here, at least twice as many subjects as factor levels seemed to be required for convergence rates higher than .9, and even much larger sample sizes for the log-normally distributed data.

In sum, fitting the more complex covariance structures with PROC MIXED can be a problem when sample sizes are small. In most cases, however, the convergence rates are very high. Note that our results on Type I error rates when selecting the model covariance structure via AIC or BIC indirectly consider problems with convergence, because if one PROC MIXED variant fails to converge, another model covariance structure will be selected. Thus, if Figs. 1–3 show that the Type I error rate for model selection via AIC or BIC is acceptable, then there were no problems with convergence.



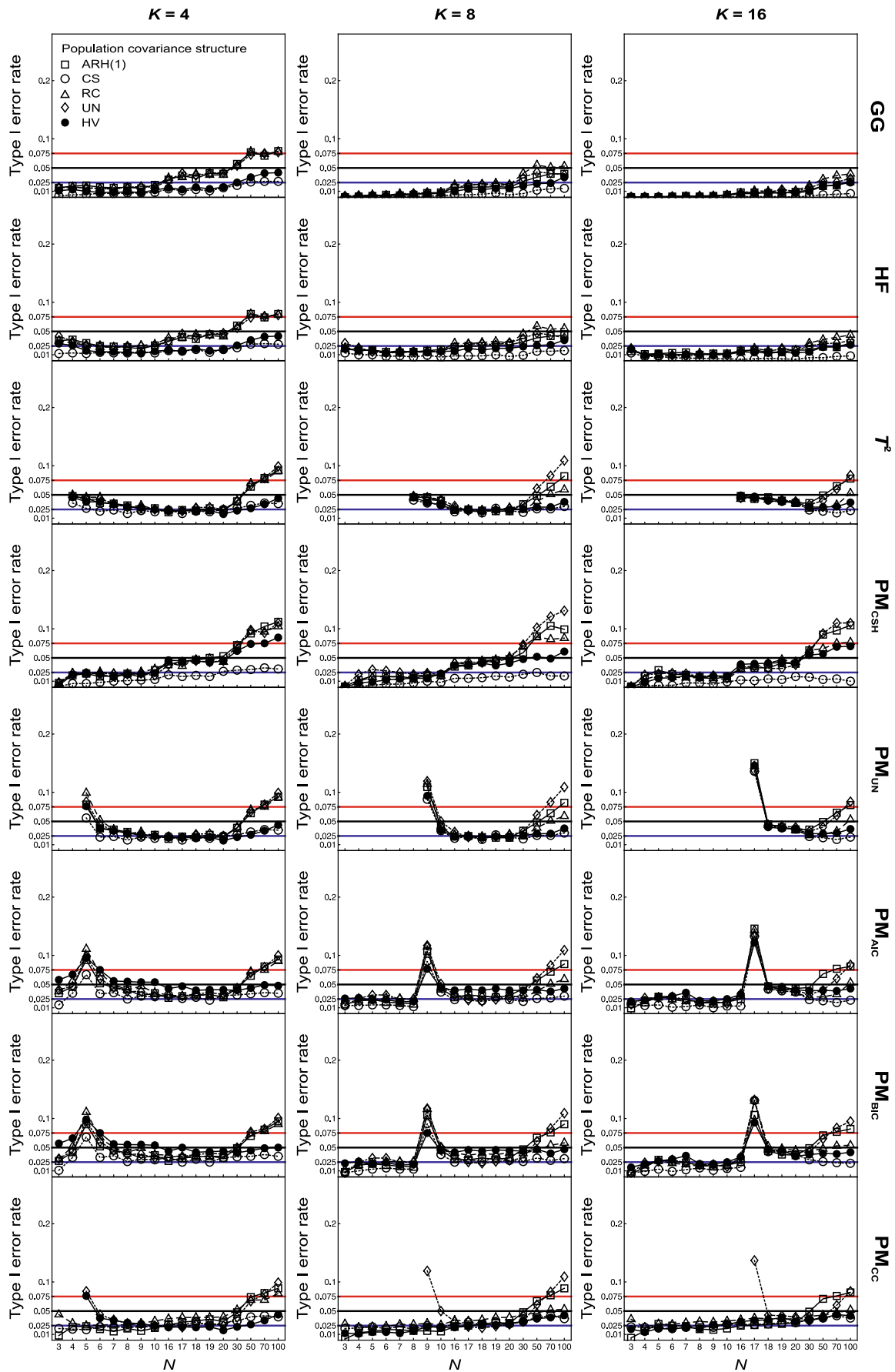


Fig. 3 Log-normally distributed data: Empirical Type I error rates for the eight analysis procedures (rows), as a function the number of factor levels (columns), sample size (N), and the population variance–covariance structure (symbols). The format is the same as in Fig. 1

Discussion

In a simulation study, we obtained the empirical Type I error rates of different procedures for analyzing data from a repeated measures design with a single within-subjects factor and no grouping factor, with a focus on extremely small sample sizes and nonnormal data. In fact, we studied for the first time the behavior of several different analysis approaches for sample sizes smaller than six combined with a high number of levels of the within-subjects factor ($K = 8$ or 16). We also studied two distributions of the response variable showing stronger deviations from normality than did the nonnormal distributions included in some previous simulation studies.

Figure 4 provides a visualization of the robustness of the different analysis approaches for the case of normal data. The vertical axis displays the number of liberal or conservative Type I error rates (according to Bradley’s, 1978, liberal criterion) across the five population covariance structures (i.e., including the nonspherical structures with homogeneous variances) and as a function of the factors varied in the simulation. The ideal procedure would show a value of zero in all conditions. The worst possible outcome would be a value of five in all conditions, which would indicate that a given analysis approach produced nonrobust Type I error

rates for all combinations of number of factor levels, sample size, and population covariance structure. The superior performance of the HF and the T^2 approaches is obvious: For these procedures, virtually all data points are located within the gray area indicating control of the Type I error rate. In contrast, for nonnormal data, Figs. 5 and 6 show that in most cases the number of nonrobust Type I error rates was higher than 0, visualizing our conclusion that none of the procedures is robust against the (admittedly rather strong) deviations from nonnormality that we studied.

We begin our discussion with the case of normal data. Here, as expected, the exact multivariate test (T^2) controlled the Type I error rate in all conditions (according to Bradley’s liberal criterion). This approach requires $N \geq K$, however, so it cannot be used in typical experiments from, for example, psychophysics, where the number of subjects is often smaller than the number of factor levels. With only very few exceptions, the univariate approach with Huynh–Feldt correction for the degrees of freedom controlled the Type I error rate even at $N = 3$, and can thus be recommended for designs with extremely small sample sizes, given that the response variable is normally distributed. Note that we only analyzed Type I error rates but did not compare the power of the T^2 and HF procedures. This is an important task for future research. Previous results for higher sample sizes have suggested that the relative power of the two approaches depends on the combination of N and K and on the deviation from sphericity of the population covariance matrix (Algina & Keselman, 1997). While the PROC MIXED variants fitting a

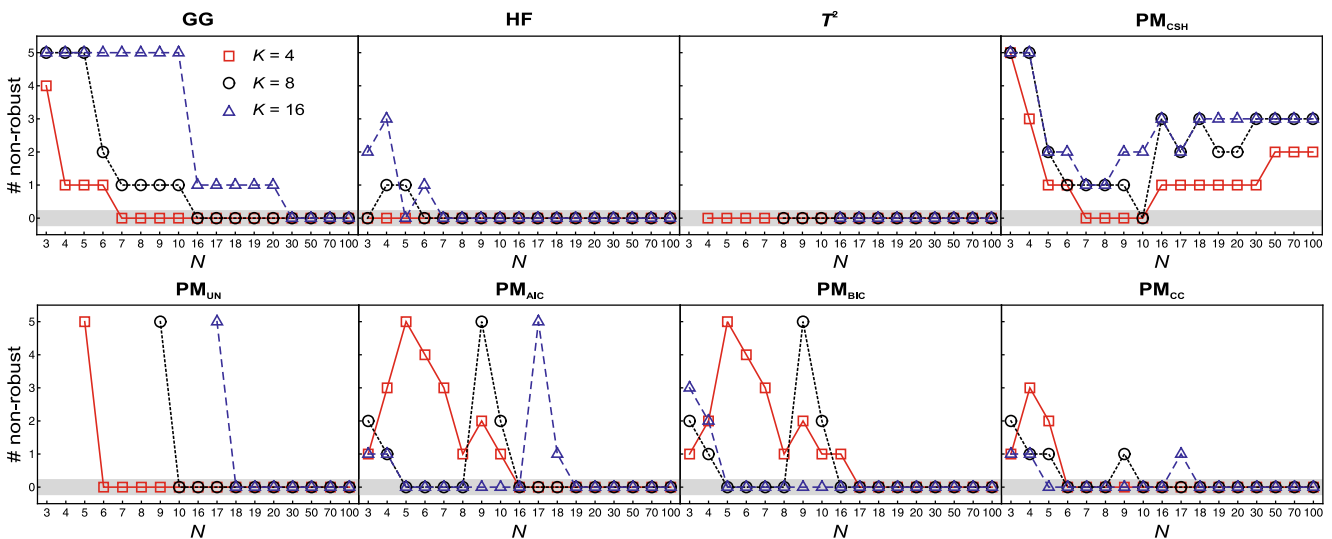


Fig. 4 Normal data: Number of nonrobust Type I error rates (according to Bradley’s liberal criterion) across the five population covariance structures, as a function of the analysis procedure (panels), number of factor levels (K), and sample size (N). Data points within the gray area indicate that the procedure controlled the Type I error rate (i.e., produced no nonrobust error rates). Red boxes indicate $K = 4$; black circles, $K = 8$; and blue triangles, $K = 16$. GG and HF indicate

univariate approaches with Greenhouse–Geisser and Huynh–Feldt df corrections, respectively; T^2 , the multivariate approach; PM_{CSH} and PM_{UN} , PROC MIXED fitting a CSH or a UN covariance structure, respectively; PM_{AIC} and PM_{BIC} , PROC MIXED with model covariance structures selected via information criteria; PM_{CC} , PROC MIXED fitting the correct (population) covariance structure

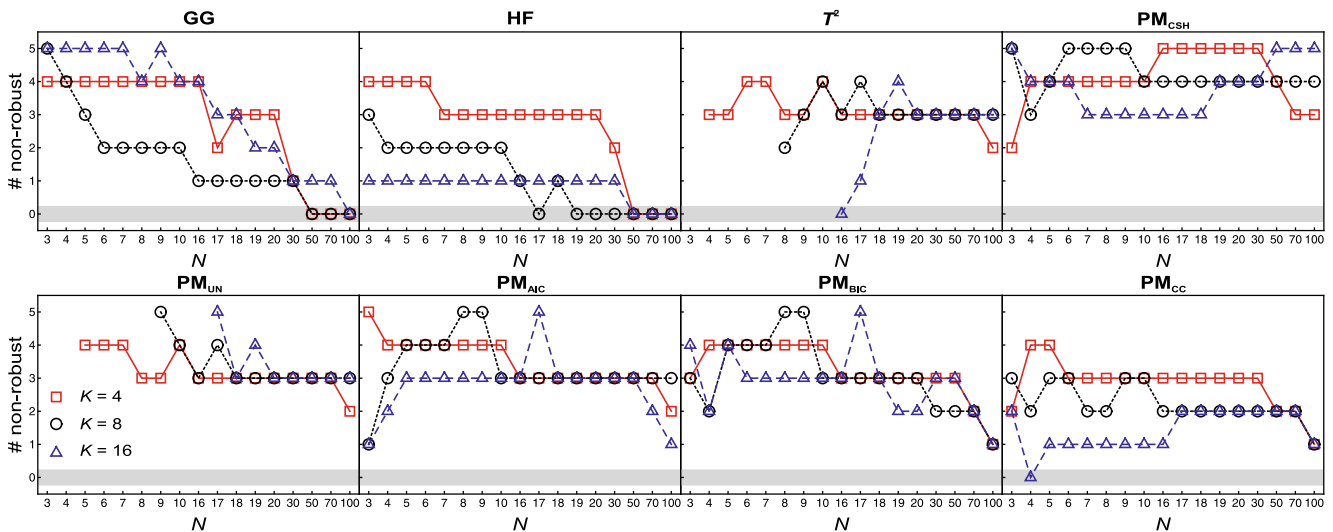


Fig. 5 Chi-square-distributed data: Number of nonrobust Type I error rates (according to Bradl liberal criterion) across the five population covariance structures. The format is the same as in Fig. 4

UN model covariance structure or selecting the model covariance structure via information criteria (AIC and BIC) controlled the Type I error rates in most conditions for normal data, they performed less convincingly than the two traditional approaches. Therefore, we do not recommend the use of PROC MIXED unless there are missing data, which should not be the case for controlled experiments. In any case, it can be concluded that several procedures for the analysis of repeated measures data show excellent control of the Type I error rate for normal data, even for extremely small sample sizes.

In contrast, for nonnormally distributed data, our data indicate serious problems with the control of Type I error rates for all analysis procedures, which is compatible with

previous studies simulating small sample sizes (see the introduction). This is visualized in Figs. 5 and 6, where most data points are located outside the gray area that indicates robust Type I error rates.

For $\chi^2(2)$ -distributed data, we observed a marked difference between the spherical (CS) population covariance structure and the nonspherical structures. For the CS population covariance matrix, error rates tended to be conservative, while for the remaining structures, most values were liberal (see Fig. 2). For the nonspherical covariance structures with homogeneous variances, the results were in between those in the two former cases. It would be interesting to study the effects of nonsphericity versus variance heterogeneity in greater detail. For log-normally distributed data,

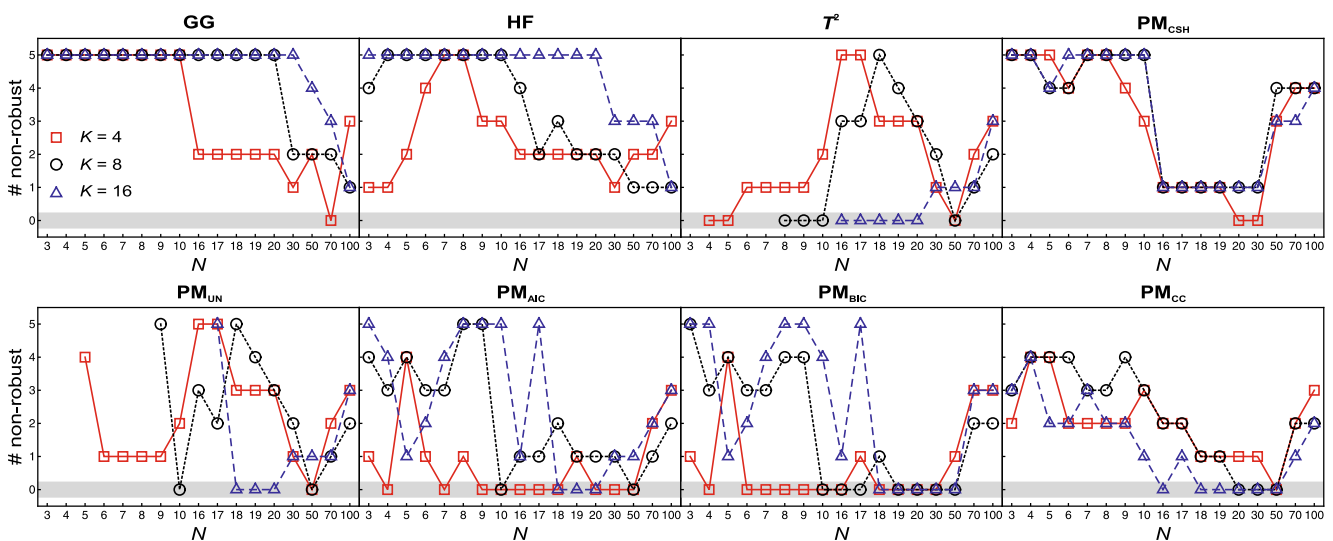


Fig. 6 Log-normally distributed data: Number of nonrobust Type I error rates (according to Bradley’s liberal criterion) across the five population covariance structures. The format is the same as in Fig. 4

we observed conservative Type I error rates in the majority of conditions, especially at smaller sample sizes. At sample sizes greater than 50, several liberal error rates were observed. For both distributions, not a single procedure showed acceptable control of Type I error rates across a wider range of settings, so that, unfortunately, none of the procedures can be recommended.

It is interesting to compare our results to the effects of nonnormality in a completely randomized (independent-groups) design containing only between-subjects factors. In that case, many studies have shown that for ANOVAs and t tests, departures from normality are not critical (e.g., Glass, Peckham, & Sanders, 1972; Harwell, Rubinstein, Hayes, & Olds, 1992; Lix, Keselman, & Keselman, 1996; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010). For t tests, the Type I errors still meet Bradley's liberal criterion even with extreme deviations from normality, at least if the sample size is higher than about 20, if there are (approximately) equal numbers of observations per group (balanced design) and if two-tailed tests are conducted (Kubinger, Rasch, & Moder, 2009; Sawilowsky & Blair, 1992). Therefore, for between-subjects designs, textbooks do not recommend against the application of the general linear model for nonnormal data (e.g., Maxwell & Delaney, 2004, pp. 111ff). For between-subjects designs, the standard ANOVA is also known to be robust to moderate violations of the homogeneity of variances, unless the sample sizes are not identical between groups (Lix et al., 1996; Maxwell & Delaney, 2004). Our results show that—unfortunately—the relatively high robustness against nonnormality of the general linear model reported for a completely randomized design does not generalize to a repeated measures design.

One of the most surprising findings was that for nonnormal correlated data, some analysis approaches produce liberal or conservative Type I error rates even at samples sizes as high as 100 (see Figs. 5 and 6). Thus, simply increasing the sample size does not guarantee robustness when the data are nonnormal. In a sense, this indicates that the intuition of experimenters from the field of psychophysics is correct: These experiments typically use only a small number of subjects but collect many trials per subject and experimental condition, rather than testing a large number of subjects but collecting only few trials. If many trials are available per subject and condition (e.g., $m = 100$), then according to the central limit theorem (see Le Cam, 1986, for a historical review) the mean of the observations across trials will be approximately normally distributed, even if the underlying population distribution of the response measure is nonnormal. Textbooks on statistics typically state that the sample mean will be approximately normally distributed if the number of observations is greater than 30 (e.g., Hays, 1988). For the log-normal distribution that we studied, the skewness and kurtosis of the sample mean with $m = 30$ observations are 1.1 and 6.0,

respectively, which indicates a distribution much closer to the normal distribution than the underlying distribution of the response measure. It would be interesting to investigate for which values of skewness and kurtosis the analysis procedures produce robust Type I error rates. Given this information, and some information concerning the distribution of the response measure, it would then be possible to decide how many trials per condition are required.

Besides averaging across a rather large number of trials per subject and experimental condition in order to make use of the central limit theorem, there are four other potential solutions to the problem of nonrobust Type I error rates for nonnormally distributed data from a repeated measures design.

First, the data values could be transformed prior to conducting the statistical tests. For a log-normal distribution, or for an approximately log-normal distribution such as response times, taking the logarithm of the observed values would be the obvious choice. In our experience, researchers analyzing response times seem reluctant to apply this transformation. One reason might be that the transformation would change the interpretation of interaction effects, which are often important in factorial designs. Several transformations are discussed or recommended in textbooks on statistics, such as, for example, an arcsine-square root transform for proportions (e.g., Maxwell & Delaney, 2004; Winer, Brown, & Michels, 1991). However, simulation studies investigating the effects of transformations for repeated measures analyses are missing, and even for data from completely randomized designs, a debate has concerned whether or not transformations should be used (Games, 1983, 1984; Levine & Dunlap, 1982, 1983). It is also not always the case that researchers will know what distribution their data will exhibit. However, at least for a completely randomized design, it might be possible to select the appropriate transform on the basis of the sample characteristics (Rasmussen, 1989). Again, no corresponding simulation studies seem to exist for repeated measures designs.

Second, the parametric procedures designed for the analysis of normal repeated measures data can be combined with “robust” estimators like trimmed means (Berkovits et al., 2000; Keselman et al., 2000; Wilcox et al., 2000) or M estimators (Wilcox & Keselman, 2003). The existing simulation studies evaluating the Type I error control of this approach do not cover a range of sample sizes, numbers of factor levels, and population covariance structures comparable to the present study, and they also indicate that the use of robust estimators will not in all cases result in sufficient control of the Type I error rate.

Third, statistical software is increasingly available for fitting generalized linear models or nonlinear models that are applicable to the data from repeated measures designs (e.g., Breslow & Clayton, 1993; Davidian & Giltinan, 1998; Hu, Goldberg, Hedeker, Flay, & Pentz, 1998; Jaeger, 2008;

Lee & Nelder, 2001; McCullagh & Nelder, 1989; Vonesh & Chinchilli, 1997). If the distribution of the response variable is known from previous research or can be estimated from the sample, then a model designed for the specific distribution can be used. For instance, for log-normal data, software including SAS PROC NLMIXED is available for fitting nonlinear mixed-effects models (e.g., Davidian & Giltinan, 2003; Jin, Hein, Deddens, & Hines, 2011; Littell et al., 2006). For binomially distributed data (e.g., error rates), several multiple logistic regression approaches accounting for the correlation structure of the data have been proposed (Hu et al., 1998; Kuss, 2002; Lipsitz, Kim, & Zhao, 1994; Neuhaus, Kalbfleisch, & Hauck, 1991; Pendergast et al., 1996; Spiess & Hamerle, 2000). However, not much is known about how these approaches behave if small sample sizes are combined with high numbers of factor levels and with different types of population covariance matrices (e.g., Austin, 2010). Specifying the correct model can also present a challenge to researchers, due to either the high flexibility or the restrictions of the software.

Finally, nonparametric approaches might be an alternative, such as, for example, the framework based on rank-score tests proposed by Brunner and colleagues (e.g., Brunner, Munzel, & Puri, 1999; Brunner & Puri, 2001). This method is available in SAS PROC MIXED, but it requires $N \geq K$ and has not been tested extensively in simulation studies.

Therefore, it has to be concluded that while there are several promising alternatives to parametric approaches assuming normality, it remains for future research to show whether these procedures will indeed solve the problems with nonrobust Type I error rates for nonnormal data in a repeated measures design that we have identified in our study.

Summary and recommendations

In this study, we obtained empirical Type I error rates for several procedures available for the analysis of data from a repeated measures design. We simulated a design with a single within-subjects factor and no grouping factors. Our focus was on specific designs often encountered in experimental psychology and the neurosciences, where high numbers of factor levels of the within-subjects factor(s) are studied in small samples, and several important response measures, such as response times, are nonnormally distributed.

Several analysis approaches showed good control of the Type I error rate for normal data, while none of the procedures was found to be robust against nonnormality.

In the following summary, we propose some guidelines for selecting an analysis procedure for completely within-subjects designs.

Figures 4–6 visualize the control of the Type I error rate for the eight analysis procedures that we studied. Each data

point shows the number of nonrobust Type I error rates at a nominal α level of .05, according to Bradley's (1978) liberal criterion, which considers an empirical Type I error rate $\hat{\alpha}$ as being acceptable if it is contained within the interval $.025 \leq \hat{\alpha} \leq .075$. While researchers often have some idea about whether their response measure is approximately normally or nonnormally distributed, it is rather unlikely that reliable information about the population variance–covariance structure will be available. Therefore, we studied five quite different population variance–covariance structures. Each data point in Figs. 4–6 shows the number of nonrobust Type I error rates across the five population covariance structures. If this number is 0 (i.e., the data point is located within the gray area), our study indicates that for a given number of factor levels (K) and sample size (N), the procedure adequately controls the Type I error rate across a wide range of possible population covariance structures.

Our simulation results show that it is primarily important to distinguish between the cases of normally and nonnormally distributed response measures.

For *normal data*, Fig. 4 shows that the multivariate approach (T^2) controls the Type I error rate. However, this procedure requires $N \geq K$, and therefore cannot be used if a small sample is studied under a high number of factor levels. For $N < K$, the univariate approach with Huynh–Feldt correction for the degrees of freedom can be recommended, which produced only very few nonrobust error rates. PROC MIXED unconditionally fitting a univariate covariance structure performed identically to the multivariate approach for $N > K + 1$. It could therefore be used in the case of missing data (MCAR or MAR), but missing data are typically not a problem in controlled laboratory experiments. The remaining procedures (GG and PROC MIXED with selection of the model covariance structure via information criteria) did not show acceptable control of Type I error rates across conditions. Note that PM_{CC} is only of theoretical interest here, because the population covariance structure will typically be unknown. Finally, our study only obtained Type I error rates. Previous studies have shown that in terms of statistical power, the recommended procedures HF and T^2 can differ quite substantially, depending on the sample size, K , and the nonsphericity of the covariance matrix. Simple rules for deciding between the two procedures can be found on page 215 in Algina and Keselman (1997), albeit only for sample sizes greater than $K + 4$.

For *nonnormally* distributed response measures, Figs. 5 and 6 show that none of the procedures that we studied was able to control the Type I error rate across a larger range of conditions. Notably, even at a sample size of 50 or 100, liberal or conservative error rates were observed in a substantial number of cases. Although several potential solutions exist to the problem of nonnormal data in a repeated measures design (see the Discussion section), it is currently

unclear which of these alternative procedures can be recommended. Therefore, our recommendation for researchers studying a nonnormal response variable is simply to collect a higher number of trials per subject and experimental condition (i.e., level of the within-subjects factor), and then for each experimental condition and subject compute the mean across trials. According to the central limit theorem, with increasing numbers of trials, this sample mean will approach a normal distribution, even if the underlying response measure strongly deviates from normality. Therefore, if a sufficient number of trials is collected per subject and experimental condition, the dependent variable in the repeated measures ANOVA can always be considered normally distributed, and therefore either the HF or the T^2 procedure can be used. The necessary number of trials per condition will of course depend on the deviation from normality of the “raw” response measure. Additional research is necessary for providing exact guidelines concerning the minimum number of trials for which robust Type I error rates can be obtained with non-normal response measures. At present, the number of 30 trials typically recommended in statistics textbooks (Hays, 1988) could be used as a lower limit.

Author Note T.F. is now at the German Federal Statistical Office. This work was supported by a grant from Johannes Gutenberg-Universität Mainz (“Förderstufe I”) to the first author. We are grateful to Marieke Hölle for helping with the figure and table preparation.

References

- Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control, AC-19*, 716–723. doi:10.1109/TAC.1974.1100705
- Algina, J., & Keselman, H. J. (1997). Detecting repeated measures effects with univariate and multivariate statistics. *Psychological Methods, 2*, 208–218.
- Algina, J., & Keselman, H. J. (1998). A power comparison of the Welch–James and improved general approximation tests in the split-plot design. *Journal of Educational and Behavioral Statistics, 23*, 152–169.
- Algina, J., & Oshima, T. C. (1994). Type I error rates for Huynh’s general approximation and improved general approximation tests. *British Journal of Mathematical and Statistical Psychology, 47*, 151–165.
- Arnau, J., Bono, R., & Vallejo, G. (2009). Analyzing small samples of repeated measures data with the mixed-model adjusted F test. *Communications in Statistics: Simulation and Computation, 38*, 1083–1103. doi:10.1080/03610910902785746
- Austin, P. J. (2010). Estimating multilevel logistic regression models when the number of clusters is low: A comparison of different statistical software procedures. *International Journal of Biostatistics, 6*(1), 16. doi:10.2202/1557-4679.1195
- Berkovits, I., Hancock, G. R., & Nevitt, J. (2000). Bootstrap resampling approaches for repeated measure designs: Relative robustness to sphericity and normality violations. *Educational and Psychological Measurement, 60*, 877–892.
- Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics, 25*, 484–498.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144–152.
- Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*, 9–25.
- Brett, M., Johnsrude, I. S., & Owen, A. M. (2002). The problem of functional localization in the human brain. *Nature Reviews Neuroscience, 3*, 243–249. doi:10.1038/Nrn756
- Brunner, E., Munzel, U., & Puri, M. L. (1999). Rank-score tests in factorial designs with repeated measures. *Journal of Multivariate Analysis, 70*, 286–317.
- Brunner, E., & Puri, M. L. (2001). Nonparametric methods in factorial designs. *Statistical Papers, 42*, 1–52.
- Benoit, C. (1924). Note sur une méthode de résolution des équations normales provenant de l’application de la méthode des moindres carrés à un système d’équations linéaires en nombre inférieur à celui des inconnues—Application de la méthode à la résolution d’un système défini d’équations linéaires (Procédé du Commandant Cholesky). *Bulletin Géodésique, 2*, 67–77. doi:10.1007/BF03031308
- Davidian, M., & Giltinan, D. M. (1998). *Nonlinear models for repeated measurement data*. Boca Raton, FL: Chapman & Hall/CRC.
- Davidian, M., & Giltinan, D. M. (2003). Nonlinear models for repeated measurement data: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics, 8*, 387–419. doi:10.1198/1085711032697
- DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*, 292–307. doi:10.1037/1082-989x.2.3.292
- Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96*, 433–458. doi:10.1037/0033-295X.96.3.433
- Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16*, 143–149. doi:10.3758/BF03203267
- Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*, 429–433. doi:10.1038/415429a
- Fleishman, A. I. (1978). Method for simulating non-normal distributions. *Psychometrika, 43*, 521–532.
- Florentine, M., Buus, S., & Poulsen, T. (1996). Temporal integration of loudness as a function of level. *Journal of the Acoustical Society of America, 99*, 1633–1644.
- Fouladi, R. T., & Shieh, Y. Y. (2004). A comparison of two general approaches to mixed model longitudinal analyses under small sample size conditions. *Communications in Statistics: Simulation and Computation, 33*, 807–824.
- Games, P. A. (1983). Curvilinear transformations of the dependent variable. *Psychological Bulletin, 93*, 382–387. doi:10.1037/0033-2909.93.2.382
- Games, P. A. (1984). Data transformations, power, and skew: A rebuttal to Levine and Dunlap. *Psychological Bulletin, 95*, 345–347. doi:10.1037/0033-2909.95.2.345
- Geisser, S., & Greenhouse, S. W. (1958). An extension of Box’s results on the use of the F distribution in multivariate analysis. *Annals of Mathematical Statistics, 29*, 885–891.
- Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. *Review of Educational Research, 42*, 237–288. doi:10.3102/00346543042003237
- Gomez, E. V., Schaalje, G. B., & Fellingham, G. W. (2005). Performance of the Kenward–Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics: Simulation and Computation, 34*, 377–392.

- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.
- Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, 24, 95–112. doi:10.1007/BF02289823
- Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte-Carlo results in methodological research: The one-factor and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, 17, 315–339. doi:10.3102/10769986017004315
- Hays, W. L. (1988). *Statistics* (4th ed.). Fort Worth, TX: Holt, Rinehart & Winston.
- Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics & Data Analysis*, 40, 685–711.
- Headrick, T. C., & Kowalchuk, R. K. (2007). The power method transformation: Its probability density function, distribution function, and its further use for fitting data. *Journal of Statistical Computation and Simulation*, 77, 229–249. doi:10.1080/10629360600605065
- Headrick, T. C., Kowalchuk, R. K., & Sheng, Y. (2010). Parametric probability densities and distribution functions for Tukey g- and -h transformations and their use for fitting data. *Applied Mathematical Sciences*, 2, 449–462.
- Headrick, T. C., Sheng, Y. Y., & Hodis, F. A. (2007). Numerical computing and graphics for the power method transformation using Mathematica. *Journal of Statistical Software*, 19(3), 1–17.
- Hearne, E. M., Clark, G. M., & Hatch, J. P. (1983). A test for serial correlation in univariate repeated-measures analysis. *Biometrics*, 39, 237–243.
- Heathcote, A., Brown, S., & Cousineau, D. (2004). QMPE: Estimating lognormal, Wald, and Weibull RT distributions with a parameter-dependent lower bound. *Behavior Research Methods, Instruments, & Computers*, 36, 277–290. doi:10.3758/bf03195574
- Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics*, 2, 360–378. doi:10.1214/aoms/117732979
- Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., & Pentz, M. A. (1998). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology*, 147, 694–703.
- Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact *F*-distributions. *Journal of the American Statistical Association*, 65, 1582–1589.
- Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, 1, 69–82. doi:10.2307/1164736
- Huynh, H., & Mandeville, G. K. (1979). Validity conditions in repeated measures designs. *Psychological Bulletin*, 86, 964–973. doi:10.1037/0033-2909.86.5.964
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59, 434–446. doi:10.1016/j.jml.2007.11.007
- Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42, 805–820. doi:10.2307/2530695
- Jin, Y., Hein, M. J., Deddens, J. A., & Hines, C. J. (2011). Analysis of lognormally distributed exposure data with repeated measures and values below the limit of detection using SAS. *Annals of Occupational Hygiene*, 55, 97–112. doi:10.1093/annhyg/meq061
- Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika*, 27, 179–182.
- Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17, 4302–4311.
- Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53, 983–997. doi:10.2307/2533558
- Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, 53, 2583–2595. doi:10.1016/j.csda.2008.12.013
- Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical and Statistical Psychology*, 54, 1–20.
- Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2002). A comparison of data analysis strategies for testing omnibus effects in higher-order repeated measures designs. *Multivariate Behavioral Research*, 37, 331–357.
- Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1999a). The analysis of repeated measurements: A comparison of mixed-model Satterthwaite *F* tests and a nonpooled adjusted degrees of freedom multivariate test. *Communications in Statistics: Theory and Methods*, 28, 2967–2999.
- Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1999b). A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology*, 52, 63–78.
- Keselman, H. J., Carriere, K. C., & Lix, L. M. (1993). Testing repeated-measures hypotheses when covariance matrices are heterogeneous. *Journal of Educational Statistics*, 18, 305–319.
- Keselman, H. J., Keselman, J. C., & Lix, L. M. (1995). The analysis of repeated measurements: Univariate tests, multivariate tests, or both? *British Journal of Mathematical and Statistical Psychology*, 48, 319–338.
- Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L. M., & Wilcox, R. R. (2000a). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. *British Journal of Mathematical and Statistical Psychology*, 53, 175–191.
- Keselman, H. J., Kowalchuk, R. K., & Boik, R. J. (2000b). An examination of the robustness of the empirical Bayes and other approaches for testing main and interaction effects in repeated measures designs. *British Journal of Mathematical and Statistical Psychology*, 53, 51–67.
- Keselman, H. J., Rogan, J. C., Mendoza, J. L., & Breen, L. J. (1980). Testing the validity conditions of repeated measures *F*-tests. *Psychological Bulletin*, 87, 479–481.
- 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, 224–242. doi:10.1177/0013164403260196
- Kubinger, K. D., Rasch, D., & Moder, K. (2009). Zur Legende der Voraussetzungen des *t*-Tests für unabhängige Stichproben [On the legend of the prerequisites of *t*-tests for independent samples]. *Psychologische Rundschau*, 60, 26–27. doi:10.1026/0033-3042.60.1.26
- Kuss, O. (2002, April). *How to use SAS for logistic regression with correlated data*. Paper presented at the 27th Annual SAS Users Group International Conference, Orlando, FL.
- Le Cam, L. (1986). The central limit theorem around 1935. *Statistical Science*, 1, 78–91.
- Lecoutre, B. (1991). A correction for the epsilon approximate test in repeated measures designs with two or more independent groups. *Journal of Educational Statistics*, 16, 371–372.
- Lee, Y., & Nelder, J. A. (2001). Modelling and analysing correlated non-normal data. *Statistical Modelling*, 1, 3–16. doi:10.1177/1471082x0100100102
- Levine, D. W., & Dunlap, W. P. (1982). Power of the *F* test with skewed data: Should one transform or not? *Psychological Bulletin*, 92, 272–280. doi:10.1037/0033-2909.92.1.272

- Levine, D. W., & Dunlap, W. P. (1983). Data transformation, power, and skew: A rejoinder to Games. *Psychological Bulletin*, *93*, 596–599. doi:10.1037/0033-2909.93.3.596
- Lipsitz, S. R., Kim, K., & Zhao, L. P. (1994). Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, *13*, 1149–1163.
- Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Cary, NC: SAS Institute, Inc.
- Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, *19*, 1793–1819. doi:10.1002/1097-0258(20000715)19:13<1793::aid-sim482>3.3.co;2-h
- Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research*, *66*, 579–619. doi:10.2307/1170654
- Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York, NY: Oxford University Press.
- MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*, 163–203. doi:10.1037/0033-2909.109.2.163
- Mauchly, J. W. (1940). Significance test for sphericity of n -variate normal populations. *Annals of Mathematical Statistics*, *11*, 204–209.
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London, U.K.: Chapman and Hall.
- Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1423–1442. doi:10.1037/0278-7393.22.6.1423
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166.
- Muller, K. E., Edwards, L. J., Simpson, S. L., & Taylor, D. J. (2007). Statistical tests with accurate size and power for balanced linear mixed models. *Statistics in Medicine*, *26*, 3639–3660.
- Neuhaus, J. M., Kalbfleisch, J. D., & Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, *59*, 25–35.
- Padilla, M. A., & Algina, J. (2004). Type I error rates for a one factor within-subjects design with missing values. *Journal of Modern Applied Statistical Methods*, *3*, 406–416.
- Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 58–71. doi:10.1037/a0020747
- Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., & Fisher, M. R. (1996). A survey of methods for analyzing clustered binary response data. *International Statistical Review*, *64*, 89–118.
- Potvin, P. J., & Schutz, R. W. (2000). Statistical power for the two-factor repeated measures ANOVA. *Behavior Research Methods, Instruments, & Computers*, *32*, 347–356. doi:10.3758/bf03207805
- Quintana, S. M., & Maxwell, S. E. (1994). A Monte-Carlo comparison of seven epsilon-adjustment procedures in repeated-measures designs with small sample sizes. *Journal of Educational Statistics*, *19*, 57–71.
- Rasmussen, J. L. (1989). Data transformation, Type I error rate and power. *British Journal of Mathematical and Statistical Psychology*, *42*, 203–213.
- Rouanet, H., & Lépine, D. (1970). Comparison between treatments in a repeated-measurement design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology*, *23*, 147–163.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–590.
- Sams, M., Paavilainen, P., Alho, K., & Näätänen, R. (1985). Auditory frequency discrimination and event-related potentials. *Electroencephalography and Clinical Neurophysiology*, *62*, 437–448.
- Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type-II error properties of the t test to departures from population normality. *Psychological Bulletin*, *111*, 352–360.
- Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics*, *7*, 512–524.
- Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology*, *6*, 147–151. doi:10.1027/1614-2241/a000016
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464. doi:10.1214/aos/1176344136
- Skene, S. S., & Kenward, M. G. (2010a). The analysis of very small samples of repeated measurements I: An adjusted sandwich estimator. *Statistics in Medicine*, *29*, 2825–2837. doi:10.1002/sim.4073
- Spieß, M., & Hamerle, A. (2000). A comparison of different methods for the estimation of regression models with correlated binary responses. *Computational Statistics and Data Analysis*, *33*, 439–455.
- Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, *123*, 34–80. doi:10.1037/0096-3445.123.1.34
- Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate non-normal distributions. *Psychometrika*, *48*, 465–471.
- Vallejo, G., & Livacic-Rojas, P. (2005). Comparison of two procedures for analyzing small sets of repeated measures data. *Multivariate Behavioral Research*, *40*, 179–205.
- Vallejo Seco, G., Izquierdo, M. C., Garcia, M. P. F., & Diez, R. J. H. (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*, 35–62.
- Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7*, 424–465. doi:10.3758/BF03214357
- Vonesh, E. F., & Chinchilli, V. M. (1997). *Linear and nonlinear models for the analysis of repeated measurements*. New York, NY: Dekker.
- Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). Amsterdam, The Netherlands: Elsevier/Academic Press.
- Wilcox, R. R., & Keselman, H. J. (2003). Repeated measures one-way ANOVA based on a modified one-step M -estimator. *British Journal of Mathematical and Statistical Psychology*, *56*, 15–25.
- Wilcox, R. R., Keselman, H. J., Muska, J., & Cribbie, R. (2000). Repeated measures ANOVA: Some new results on comparing trimmed means and means. *British Journal of Mathematical and Statistical Psychology*, *53*, 69–82.
- Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York, NY: McGraw-Hill.
- Wolfinger, R. D. (1996). Heterogeneous variance: Covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, *1*, 205–230.