Violation of the Sphericity Assumption and Its Effect on Type-I Error Rates in Repeated Measures ANOVA and Multi-Level Linear Models (MLM)

Nicolas Haverkamp and *André Beauducel*

Department of Arts and Humanities, Institute of Psychology, University of Bonn, Bonn, Germany

We investigated the effects of violations of the sphericity assumption on Type I error rates for different methodical approaches of repeated measures analysis using a simulation approach. In contrast to previous simulation studies on this topic, up to nine measurement occasions were considered. Effects of the level of inter-correlations between measurement occasions on Type I error rates were considered for the first time. Two populations with non-violation of the sphericity assumption, one with uncorrelated measurement occasions and one with moderately correlated measurement occasions, were generated. One population with violation of the sphericity assumption combines uncorrelated with highly correlated measurement occasions. A second population with violation of the sphericity assumption combines moderately correlated and highly correlated measurement occasions. From these four populations without any between-group effect or within-subject effect 5,000 random samples were drawn. Finally, the mean Type I error rates for Multilevel linear models (MLM) with an unstructured covariance matrix (MLM-UN), MLM with compound-symmetry (MLM-CS) and for repeated measures analysis of variance (rANOVA) models (without correction, with Greenhouse-Geisser-correction, and Huynh-Feldt-correction) were computed. To examine the effect of both the sample size and the number of measurement occasions, sample sizes of $n = 20, 40, 60, 80,$ and 100 were considered as well as measurement occasions of $m = 3, 6,$ and 9. With respect to rANOVA, the results plead for a use of rANOVA with Huynh-Feldt-correction, especially when the sphericity assumption is violated, the sample size is rather small and the number of measurement occasions is large. For MLM-UN, the results illustrate a massive progressive bias for small sample sizes ($n = 20$) and $m = 6$ or more measurement occasions. This effect could not be found in previous simulation studies with a smaller number of measurement occasions. The proportionality of bias and number of measurement occasions should be considered when MLM-UN is used. The good news is that this proportionality can be compensated by means of large sample sizes. Accordingly, MLM-UN can be recommended even for small sample sizes for about three measurement occasions and for large sample sizes for about nine measurement occasions.

Keywords: multilevel linear models, mixed linear models, hierarchical linear models, repeated measures ANOVA, simulation study

INTRODUCTION

Multilevel linear models (MLM) have been discussed as an alternative to repeated measures analysis of variance (rANOVA; Gueorguieva and Krystal, 2004; Arnau et al., 2010; Goedert et al., 2013) and, sometimes, researchers have even been urged to use MLM instead of rANOVA (Boisgontier and Cheval, 2016). Although MLM are increasingly used instead of rANOVA, the terminology is heterogeneous in this area. Different labels are used to denote models encompassing fixed and random effects, covariance pattern models, and regression models that are based on more than one data level, where the levels are typically defined by the measurement occasions nested within individuals. These models are referred to as hierarchical linear models (Bryk and Raudenbush, 1992; Raudenbush and Bryk, 2002), as (general) mixed linear models (McLean et al., 1991; Arnau et al., 2010), mixed effects models (Gueorguieva and Krystal, 2004), multilevel linear models (Hox, 2002), or as multilevel models (Maas and Hox, 2005; Lucas, 2014). We will follow Tabachnick and Fidell (2013) in using MLM to denote these models in the following.

Here, we highlight three advantages of MLM over rANOVA: First, MLM allows to model data that correspond to a multi-level structure. Whenever researchers assume at least one level of data being nested within another level, MLM is appropriate. If there is only one level of measurement occasions and one level of individuals, rANOVA is also an appropriate model, but for any more complex structure comprising several levels, MLM will be more appropriate than rANOVA (Baayen et al., 2008). Second, MLM is robust even when there are several randomly distributed missing values. Especially in research designs with several measurement occasions, a large number of missing values can occur. Since parameters (e.g., slope parameters) are estimated for each individual, there is no requirement for complete data over occasions. A third advantage is the possibility to compare MLM with different assumptions on the covariance structure of the data. For example, models with auto-regressive covariance structure, with uncorrelated structure, or with compound-symmetry (CS) are possible. When there are no specific assumptions on the covariances, MLM with an unstructured covariance (UN) matrix can be specified. To summarize, MLM has substantial advantages when compared to rANOVA.

Given the important advantages of MLM over rANOVA, MLM will be the method of choice for several repeated measures designs. Of course, the appropriateness of MLM results also depends on data characteristics, modeling options, and estimation procedures (Tabachnick and Fidell, 2013; Lucas, 2014), so that simulation studies are necessary to ascertain the quality of MLM results. Since several aspects can be varied in MLM, a single simulation study cannot cover all relevant aspects, so that different studies with a focus on different aspects have been performed. With respect to the estimation method (e.g., maximum likelihood vs. restricted maximum likelihood), some simulation studies indicate that restricted maximum likelihood is more accurate (Maas and Hox, 2005), but it seems that restricted maximum likelihood is superior for random effects and not necessarily for fixed effects (West et al., 2007). Further results

and recommendations are available with regard to sample size, group size, and number of groups. However, Maas and Hox (2005) noted that some inconsistencies of their results on sample size with the results of other simulation studies were probably related to the use of different simulation designs and different simulated conditions. The multiple options that are possible with MLM, the flexibility of the method, may have enhanced the specificity of results. For example, the statistical power and Type I error rate of MLM based on autoregressive covariances and based on unstructured covariances has been investigated by Gueorguieva and Krystal (2004). They found the lowest Type I error rate for the MLM based on autoregressive covariances. It should, however, be noted that their simulated data had an autoregressive structure. This indicates that an optimal Type I error control results when the covariance structure specified in MLM corresponds to the empirical covariance structure. In contrast, Kowalchuk et al. (2004) found that MLM based on UN performed similarly to fitting the true covariance structure and, under certain conditions, showed better Type I error control. Thus, the results of simulation studies on the specification of MLM for obtaining optimal Type I error rates are not conclusive.

Goedert et al. (2013) performed a simulation study with another type of data containing violations of the sphericity assumption and found a superior statistical power of MLM with unstructured covariances (MLM-UN) when compared to rANOVA. In their simulation study, the MLM based on the F-distribution with between-within degrees-of-freedom (West et al., 2007) showed no bias in Type I error rates, whereas MLM-UN based on Wald's z had led to large Type I error rates. On this basis, Goedert et al. (2013) recommended the use of MLM-UN based on the F-statistic instead of MLM-UN based on z or rANOVA, especially for small samples. We followed their recommendation in performing MLM-UN based on F in the following. In Goedert et al. (2013), the Type I error rates of MLM-UN based on F and rANOVA with Greenhouse-Geisser (GG) correction (Greenhouse and Geisser, 1959) were rather similar with six measurement occasions and a sample size of at least 30 cases.

Gueorguieva and Krystal (2004) also found widely acceptable and similar Type I error rates for MLM-UN and rANOVA with Greenhouse-Geisser correction (rANOVA-GG) in a simulation study based on four measurement occasions. However, in their simulation study the Type I error rates were slightly more correct for MLM based on compound-symmetry (MLM-CS). It should be noted that the CS assumption is related to the sphericity assumption of rANOVA. The CS assumption is more restrictive than the sphericity assumption (Field, 1998) so that MLM with CS will also satisfy the sphericity assumption. It would therefore be of interest to compare MLM based on CS with rANOVA results. Since rANOVA is always based on the sphericity assumption, it will also be important to compare MLM based on CS, which is more restrictive than rANOVA, with MLM based on UN, which is less restrictive than rANOVA. In order to provide a comprehensive description of the effects, a comparison of MLM and rANOVA should be based on data sets that satisfy the CS assumption as well as the sphericity assumption and it should furthermore be based on data-sets that violate both the

CS assumption and the sphericity assumption. Data that are conform to the sphericity assumption and that simultaneously violate the CS assumption are very specific (Field, 1998) and will therefore occur very rarely so that they are not interesting for a comparison of MLM with rANOVA. Since violations of the sphericity assumption are known to result in progressive Type I error rates of rANOVA, the rANOVA-GG has been proposed in order to compensate for the progressive bias. The probably less conservative correction of the progressive bias of Type I error rates of rANOVA proposed by Huynh and Feldt (1976) should also be considered.

Since the focus of Goedert et al. (2013) was laid on data relevant for research on spatial neglect, they only investigated data in which the sphericity assumption and the CS assumption were violated. The aim of the present simulation study was to extend their results on Type I error rates beyond the specific data characteristics that are relevant for research on spatial neglect. Gueorguieva and Krystal (2004) also simulated data with a mild violation of the sphericity assumption data because they investigated an autoregressive covariance structure. We will therefore investigate whether the results provided by Goedert et al. (2013) as well as Gueorguieva and Krystal (2004) can also be found for data that are conform to the sphericity assumption. Accordingly, we will compare MLM and rANOVA for data sets with and without violation of the sphericity assumption and for MLM based on UN and CS.

Simulation studies that are based on rather specific models, options, and data yield rather specific results, which might, of course, be relevant for the respective field of research when the MLM used in the simulation corresponds to the MLM that is typically used in the respective research. However, these simulation studies might be complemented with simulation studies with a focus on rather simple models and data, which do not depend so much on a large number of specific modeling options and data characteristics. Even when the focus of simulation studies will always be limited, studies that are based on very simple models and data characteristics may provide a baseline for the evaluation of more complex models. For this reason, the present simulation study will only investigate the abovementioned effects of violation vs. non-violation of the sphericity assumption on Type I error rates in MLM (based on an unstructured covariance matrix and on compound-symmetry) and rANOVA-models (without correction, with Greenhouse-Geisser-correction, and Huynh-Feldt-correction) without any between-group effect. Because rANOVA cannot be used for the simultaneous analysis of several data levels, a comparison of MLM and rANOVA does not make sense for such complex data. The current simulation study is therefore only based on a subset of repeated measures data that can be analyzed by means of rANOVA and MLM. Since the data and models analyzed in this study do not contain fixed between group effects, restricted maximum likelihood estimation will be used, since this estimation method has been shown to be most exact for random effects models (West et al., 2007). The restriction to the class of simple within-subjects models allows for an analysis of up to nine measurement occasions, which has not been done before. Kowalchuk et al. (2004) concluded

MLM-UN performs at least as well for Type I error control as MLM with known covariance matrices, which would be MLM-CS for the present data. However, their simulation study was only based on four measurement occasions. Accordingly, a central aim of the present simulation study is to investigate whether the results of Kowalchuk et al. (2004) can be generalized to more than four measurement occasions. Moreover, the abovementioned restrictions allow for 5,000 samples to be drawn from the population in each condition in order to reach substantial robustness of results. An empirical example illustrates the relevance of rANOVA and MLM for the identification of repeated measures effects.

MATERIALS AND METHODS

The simulations were performed with IBM SPSS Statistics Version 23.0.0.3. Three factors were varied systematically in the simulation study to investigate their effect on the results of repeated measures analyses: Violations of the sphericity assumption, sample sizes and number of measurement occasions. The effect of the overall inter-correlation of dependent variables was also explored. To investigate the "pure" effect of the respective analysis method on the resulting Type I error rate, neither a between-subject effect nor a within-subject effect were fixed. Thus, there was no between-group effect and the slope across measurement occasions was zero.

Concerning the sphericity assumption, two conditions were established: Under the first condition (sphericity), the sphericity assumption was not violated in the population. For each individual i there were $t = 1$ to $m = 3$, 6, or 9 measurement occasions. A population of normally distributed, z-standardized and uncorrelated variables z_{ti} ($E[z_{ti}] = 0$; $\text{Var}[z_{ti}] = 1$) was generated by means of the SPSS Mersenne Twister random number generator for this condition. Each variable represented the dependent variable measured at one measurement occasion. In the second condition (non-sphericity), the sphericity assumption was violated in the population. For this condition, a population of normally distributed, z-standardized dependent variables was generated. Again, each variable represented the dependent variable measured at one measurement occasion. In order to realize the violation of the sphericity assumption, the correlation between the dependent variables at odd measurement occasions was 0.80 and the correlation between the dependent variables at even measurement occasions was zero. Knuth (1981) described the weighting procedure for the generation of correlated variables. Accordingly, the correlated dependent variables y_{ti} were generated by means of

$$y_{ti} = \sqrt{0.80}\,c_i + \sqrt{0.20}\,z_{ti}, \quad for\ t = 1, 3, 5, 7, and\ 9, \quad (1)$$

where c_i and z_{ti} are the scores of individual i on the respective z-standardized, normal distributed random variables. For $t = 2, 4, 6$, and 8, the equation for generation of the uncorrelated dependent variables was simply $y_{ti} = z_{ti}$. A syntax example showing the generation of data corresponding to the sphericity assumption and for data with a violation of the sphericity assumption according to Equation (1) is given in the Appendix

(Supplementary Material). A similar procedure was used for the generation of dependent variables with a population inter-correlation of 0.50 without violation of the sphericity assumption. Finally, a population with violation of the sphericity assumption based on dependent variables with a population correlation of 0.50 for $t = 2, 4, 6$, and 8 and with a population correlation of 0.80 for $t = 1, 3, 5, 7$, and 9 was generated. The violation of the sphericity assumption was less pronounced in the population based on dependent variables with a correlation of 0.50 and 0.80 than in the population based on dependent variables with a correlation of 0.00 and 0.80.

The MLM for the prediction of the dependent variables is

$$y_{ti} = \gamma_{00} + U_{0i} + (\gamma_{10} + U_{1i})time_t + e_{ti}, \quad (2)$$

where γ_{00} is the grand mean intercept, U_{0i} is the individual intercept deviation, γ_{10} is the grand mean slope, U_{1i} is the individual slope deviation, $time_t$ is the level of the measurement occasion with $time_1 = 0$ to $time_9 = 8$, and e_{ti} is the deviation of individual i at measurement occasion t. The grand mean intercept γ_{00} and the grand mean slope γ_{10} were zero in the population. Although γ_{10} was zero in the population, the corresponding parameter estimate might differ significantly from zero in the sample. The corresponding parameter estimate of rANOVA might also differ significantly from zero in the sample. This is the Type I error in the present study.

The p-values of a significance test follow a continuous uniform distribution between 0 and 1 when no population effect is simulated in the data (Murdoch et al., 2008). The corresponding standard deviation is $\hat{\sigma} = \sqrt{1/12} = 0.29$ (Krishnamoorthy, 2006) and the corresponding estimate of the standard error is $SE = 0.29/\sqrt{500} = 0.013$ for $n = 500$ and $SE = 0.29/\sqrt{1,000} = 0.009$ for $n = 1,000$. Although the expectancy of p-values is 0.50 for this condition, the mean p-values will be 0.491 or below and 0.509 or greater in about one third of the simulation studies with $n = 1,000$. Accordingly, the sampling error affects the number of p-values below and above a given threshold. Since the present simulation study is based on the 0.05 alpha level, the second digit of the mean p-values should precisely be estimated so that an SE close to 0.01 might be critical. It was therefore decided to base the simulation on $n = 5,000$ per condition with a corresponding SE of 0.004 so that the second digit of the p-values is estimated with rather high precision.

Accordingly, 5,000 samples were drawn from the populations of generated variables, submitted to all of the different analysis methods and the average Type I error rate was computed for every sample under the two conditions (sphericity vs. non-sphericity) for each of the following analysis methods: Repeated measures ANOVA without correction = rANOVA, rANOVA with Greenhouse-Geisser-correction = rANOVA-GG, rANOVA with Huynh-Feldt-correction = rANOVA-HF, MLM with compound-symmetry = MLM-CS and MLM with Unstructured Covariance Matrix = MLM-UN.

Two additional factors that were considered here were the sample size and the number of measurement occasions. Sample sizes were $n = 20, 40, 60, 80$, and 100 and measurement occasions were $m = 3, 6$, and 9. Accordingly, the simulation

study comprised 150 conditions (= sphericity[2] × analysis methods[5] × n[5] × m[3]) with 5,000 samples for each condition. For each condition, the Type I error rate was reported for the 0.05 alpha-level.

To evaluate whether this simulation is able to deliver stable results, the standard deviation was computed for every average Type I error rate and included in the following graphics which illustrate the simulation findings.

RESULTS

Simulation Study

The Type I error rates for three measurement occasions and sphericity were close to what was expected at an alpha level of 0.05 and they were not substantially affected by sample size and method of data analysis (see **Figure 1A**). For the non-sphericity condition and three measurement occasions, the Type I error rates were close to the expectation for all methods except rANOVA and MLM-CS. For rANOVA and MLM-CS, a slight progressive bias was found (see **Figure 1B**). Note that here and in the following the results for rANOVA and MLM-CS were so similar that the respective lines completely overlap. Again, the effect of sample size was not substantial. As a measure of uncertainty, the standard deviations of the Type I error rates were marked as error bars in the figures. They were smaller than 0.5 percent indicating a considerable robustness of the results of the different runs of the simulation.

The Type I error rates for three measurement occasions and sphericity were similar when the dependent variables in the population were correlated: Only minimal differences occurred in terms of slightly lower Type I error rates across all methods for $n = 20$ and $n = 40$ (see **Figure 1C**). Under the condition of non-sphericity and three correlated measurement occasions, the progressive bias for rANOVA and MLM-CS was considerably smaller than for the non-correlated population (see **Figure 1D**). This was expected because the violation of the sphericity assumption was less pronounced in the population with dependent variables correlated for 0.50 and 0.80 than in the population with dependent variables correlated for 0.00 and 0.80.

Type I error rates were as expected for all methods except MLM-UN and rANOVA-GG for the sphericity condition and six measurement occasions. For MLM-UN, a strong progressive bias occurred for $n = 20$, a small progressive bias was found for $n = 40$ and a small conservative bias was found for rANOVA-GG with $n = 20$ (see **Figure 2A**). In the case of non-sphericity and six measurement occasions, there was again a small progressive bias of the Type I error rates with MLM-CS and rANOVA. Moreover, a strong progressive bias occurred for MLM-UN and $n = 20$ and a small progressive bias occurred for MLM-UN with $n = 40$ and $n = 60$ (see **Figure 2B**).

When the sphericity assumption was not violated, the Type I error rates for the population with correlated dependent variables did practically not differ from the results of the uncorrelated condition for six measurement occasions (see **Figure 2C**). Similar to the results for three measurement occasions, the progressive bias for rANOVA and MLM-CS for the correlated population

turned out to be a bit smaller under the non-sphericity condition for intercorrelations of 0.50 and 0.80 between the dependent variables than for the non-sphericity condition for intercorrelations of 0.00 and 0.80 (see **Figure 2D**). This was expected because the violation of the sphericity assumption was less pronounced in the population with dependent variables that were correlated for at least 0.50.

For the sphericity condition and nine measurement occasions, the mean Type I error rates were again as expected for all methods except MLM-UN and rANOVA-GG: A massive progressive bias was found for MLM-UN with $n = 20$ as well

as $n = 40$, while there was still substantial progressive bias for $n = 60$; small conservative biases occurred with rANOVA-GG for $n = 20$ as well as $n = 40$ (see **Figure 3A**). For the condition of non-sphericity and nine measurement occasions, there was a substantial progressive bias of the Type I error rates with MLM-CS and rANOVA. Moreover, a very strong progressive bias occurred for MLM-UN and $n = 20$ as well as for $n = 40$ and a small progressive bias occurred for MLM-UN with $n = 60$ (see **Figure 3B**). The differences between the methods that are reported here were all considerably larger than the corresponding standard deviations.
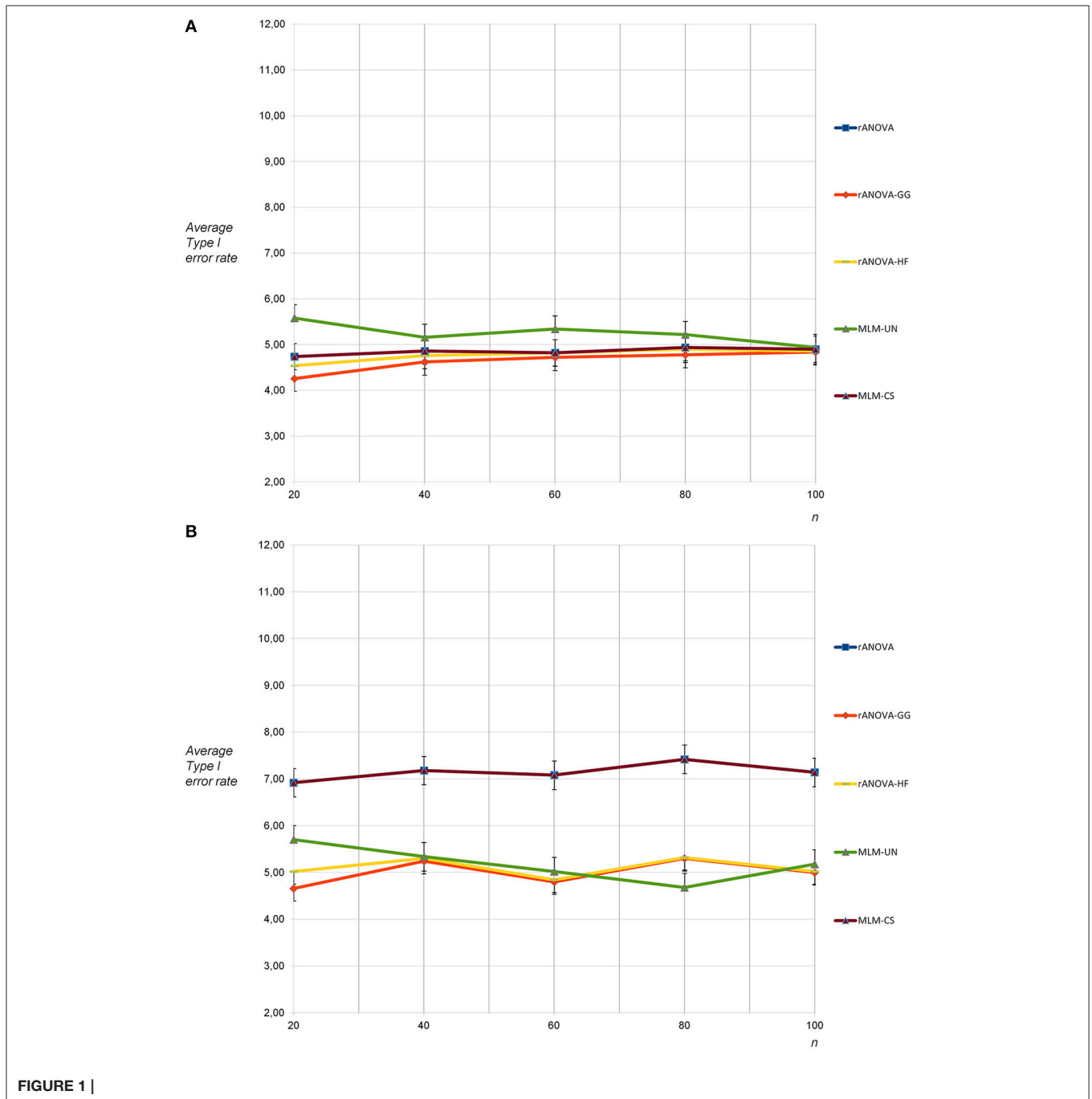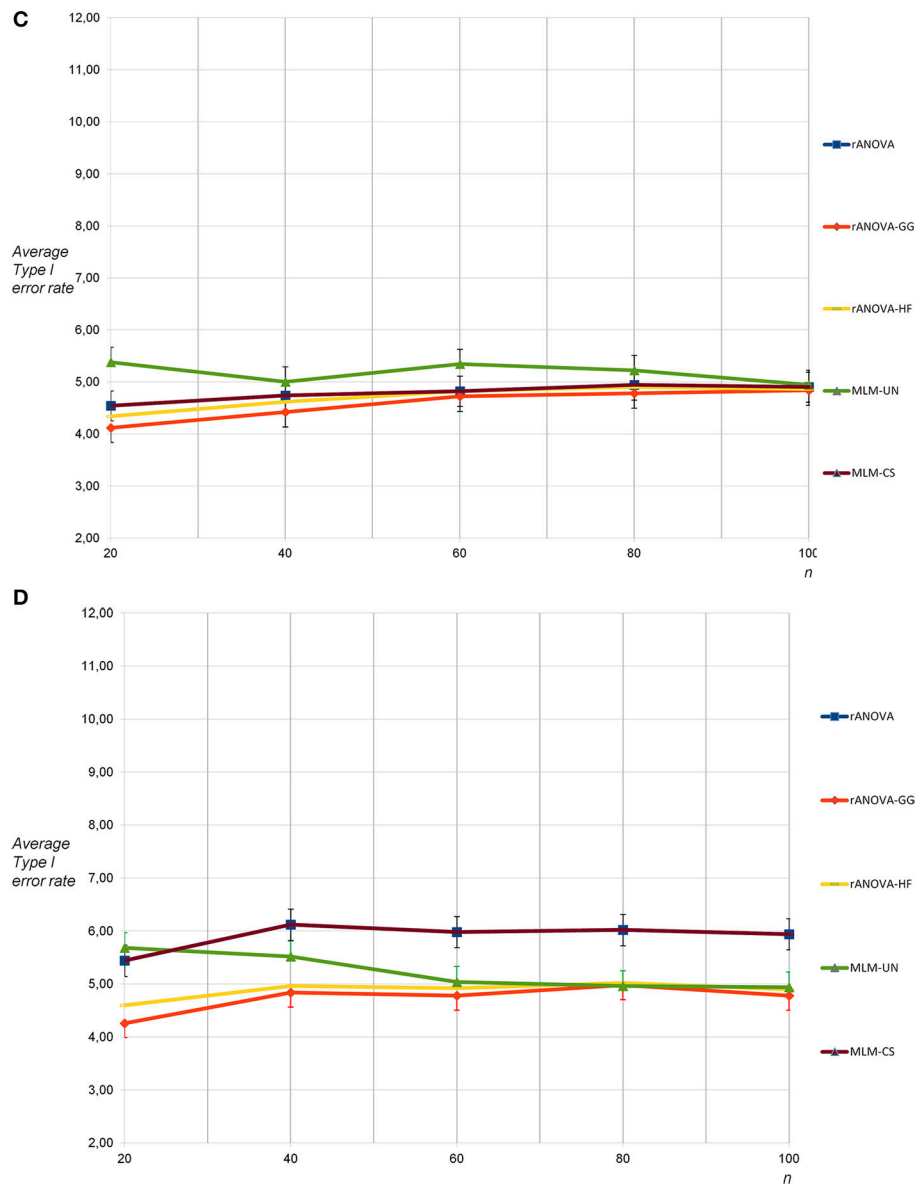

FIGURE 1 |

FIGURE 1 | (A) Average Type I error rates for 5,000 tests: no sphericity violation, three measurement occasions, uncorrelated dependent variables. **(B)** Average Type I error rates for 5,000 tests: sphericity violation, three measurement occasions, uncorrelated and highly correlated dependent variables. **(C)** Average Type I error rates for 5,000 tests: no sphericity violation, three measurement occasions, moderately correlated dependent variables. **(D)** Average Type I error rates for 5,000 tests: sphericity violation, three measurement occasions, moderately correlated and highly correlated dependent variables.

For nine measurement occasions, the results for the correlated population under the condition of sphericity showed no differences in the Type I error rates for all methods when compared to the rates for uncorrelated dependent variables (see **Figure 3C**). When the sphericity assumption was violated, the progressive bias of MLM-CS and rANOVA was smaller for the correlated dependent variables (see **Figure 3D**). This was expected because the violation of the sphericity assumption was less pronounced for this population. The remaining results for this population were similar to the results for nine measurement occasions based on populations containing a combination uncorrelated and highly correlated dependent variables.

Empirical Example

Ninety German male participants filled out nine newly developed items for the measurement of reward sensitivity (age: $M = 20.34$, $SD = 3.21$). The items were likert scaled ranging between total disagreement and complete agreement. The item means ranged between 4.08 ($SD = 1.27$) and 4.52 ($SD = 1.09$). The overall differences between item means were not significant for rANOVA [$F_{(8, 712)} = 1.61$, $p = 0.120$] and MLM-CS [$F_{(8, 712)} = 1.61$, $p = 0.120$], as well as for rANOVA-GG [$F_{(6.00, 533.68)} = 1.61$, $p = 0.144$], rANOVA-HF [$F_{(6.48, 576.52)} = 1.61$, $p = 0.137$], and MLM-UN [$F_{(8, 89)} = 1.76$, $p = 0.095$]. The Mauchley-Test [$W_{(35)} = 0.248$, $p < 0.01$] and the χ^2-difference for MLM-CS vs. MLM-UN

model fit ($\Delta\chi^2 = 141.43$, $\Delta df = 43$, $p < 0.01$) indicated that the sphericity assumption was violated in these data. However, rANOVA-GG, rANOVA-HF, and MLM-UN indicate that the overall mean differences between the items were not significant.

DISCUSSION

A simulation study for the investigation of the mean Type I error rates of different analysis methods (MLM-UN, MLM-CS,

rANOVA without correction, rANOVA-GG and r-ANOVA-HF) was performed under the conditions of violation vs. non-violation of the sphericity assumption and for sample sizes of $n = 20$, 40, 60, 80, and 100 as well as for $m = 3$, 6, and 9 measurement occasions. The simulations were based on two populations without violation of the sphericity assumption: One population with uncorrelated dependent variables and one population with correlated dependent variables. Moreover, there were two populations with violation of the sphericity assumption: One population with a combination of uncorrelated
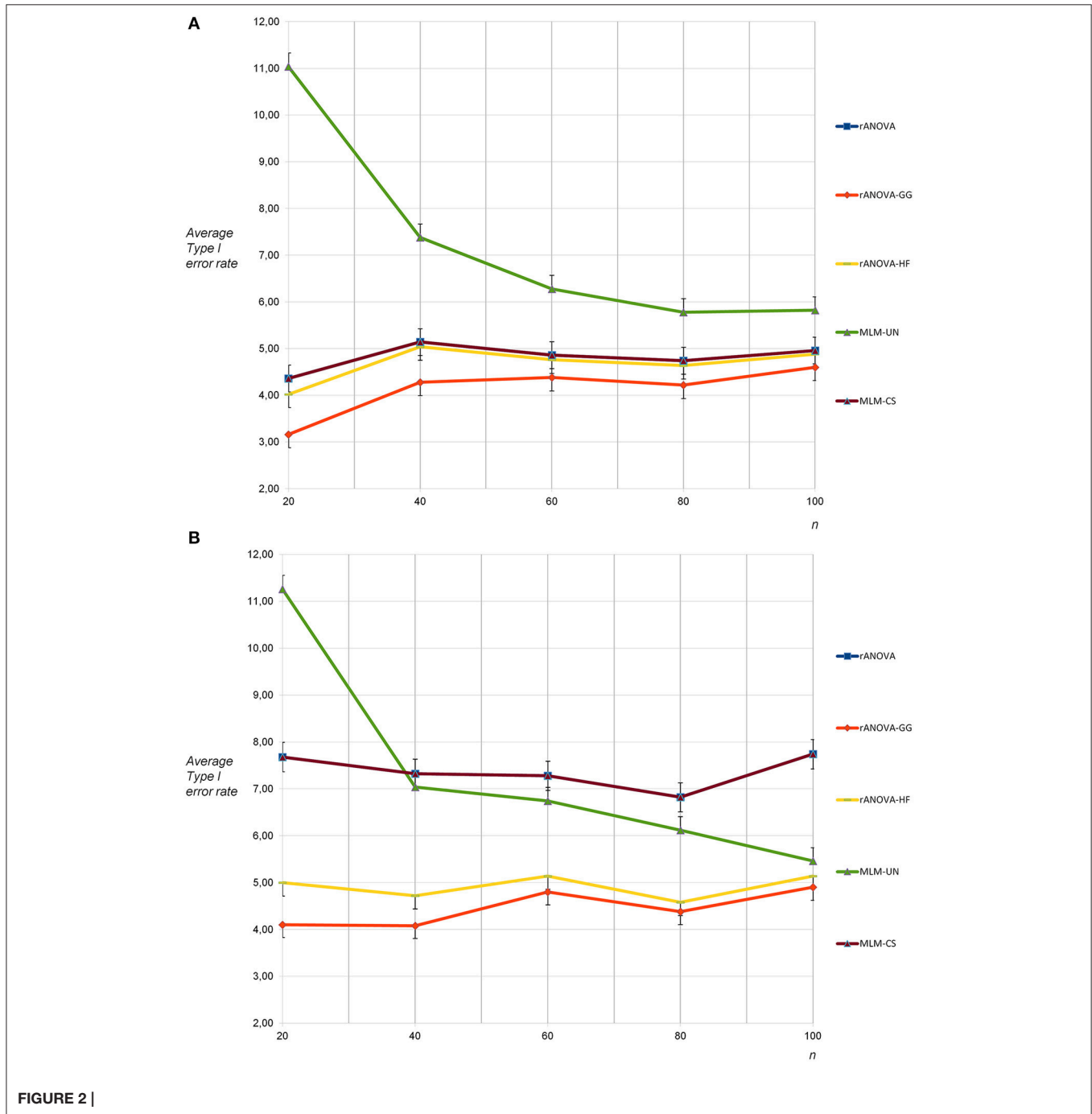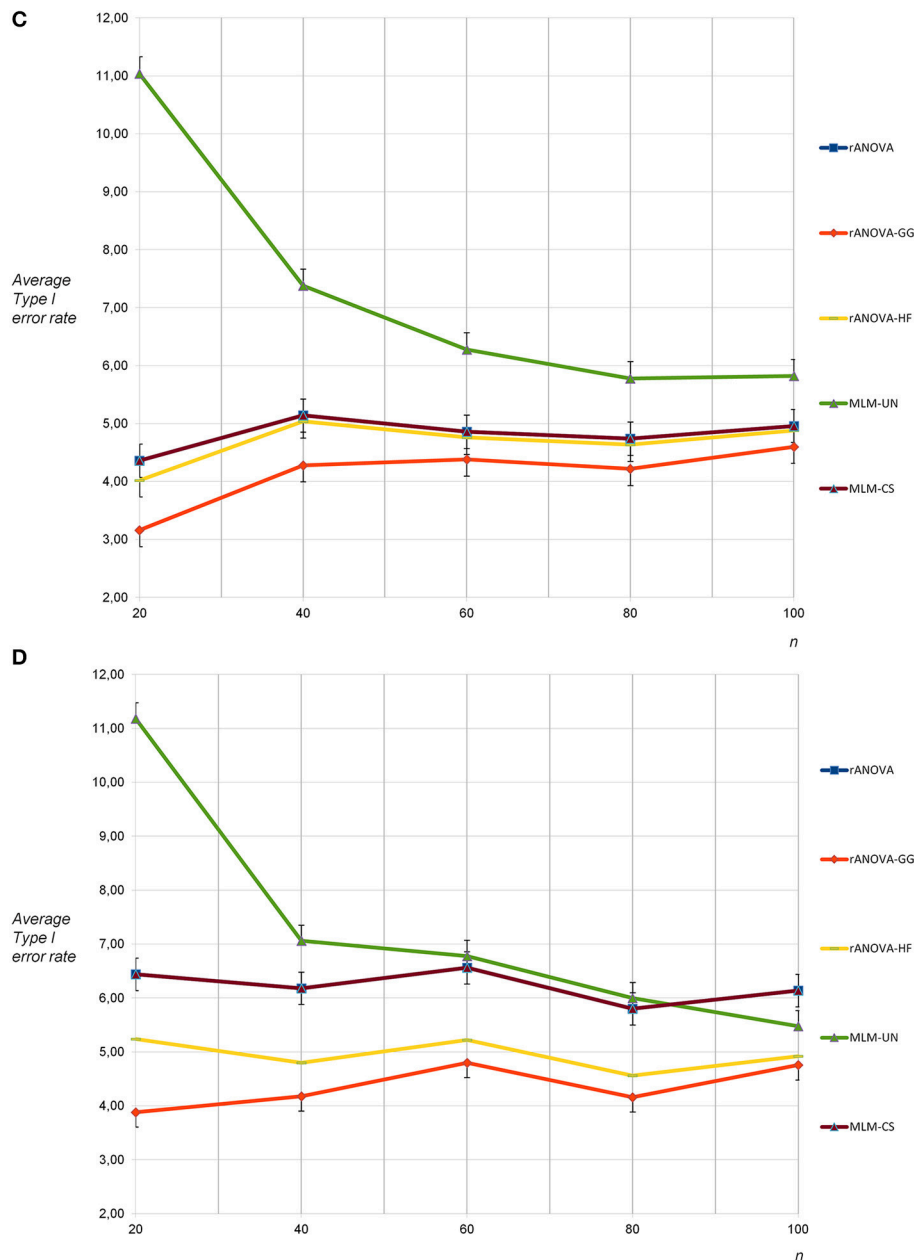

FIGURE 2 |

FIGURE 2 | (A) Average Type I error rates for 5,000 tests: no sphericity violation, six measurement occasions, uncorrelated dependent variables. **(B)** Average Type I error rates for 5,000 tests: sphericity violation, six measurement occasions, uncorrelated and highly correlated dependent variables. **(C)** Average Type I error rates for 5,000 tests: no sphericity violation, six measurement occasions, moderately correlated dependent variables. **(D)** Average Type I error rates for 5,000 tests: sphericity violation, six measurement occasions, moderately correlated and highly correlated dependent variables.

and highly correlated dependent variables and one population with moderately correlated and highly correlated dependent variables.

The simulation showed the following results: A slight progressive bias for rANOVA and MLM-CS was found in case of a violation of the sphericity assumption. This effect could be demonstrated regardless of the sample size as well as the number of measurement occasions. For MLM-UN, a massive

progressive bias for small sample sizes ($n = 20$) and $m = 6$ or more measurement occasions occurred. The progressive bias of MLM-UN was substantial for nine measurement occasions and up to medium sample sizes ($n = 60$). The mean Type I error rates for rANOVA-GG showed a small conservative bias for $m = 6$ or more measurement occasions and small sample sizes ($n = 20$) when the sphericity assumption was not violated.

The most general result of the present simulation study is that there was a substantial progressive bias of Type I error rates for MLM-UN for nine measurement occasions with sample sizes of $n = 60$ and below. This progressive bias for MLM-UN occurred when the population data was conform to the sphericity assumption but it also occurred when the sphericity assumption was violated in the population data. The progressive bias of MLM-UN for nine measurement occasions with sample sizes of $n = 60$ and below was even greater than 0.075, the upper level of Bradley's (1978) liberal criterion for the evaluation of an empirical estimate $\hat{\alpha}$ of the Type I error rate ($0.5\alpha \geq \hat{\alpha} \leq 1.5\alpha$). It should be noted that MLM was based on the F-statistic as recommended by Goedert et al. (2013) because MLM based on Wald's z already showed a progressive bias of Type I error rates in their simulation study. Thus, the present study extends Goedert et al.'s (2013) finding of a progressive bias for MLM to MLM-UN based on a large number of measurement occasions and small sample sizes, even when based on the F-statistic. Thus,
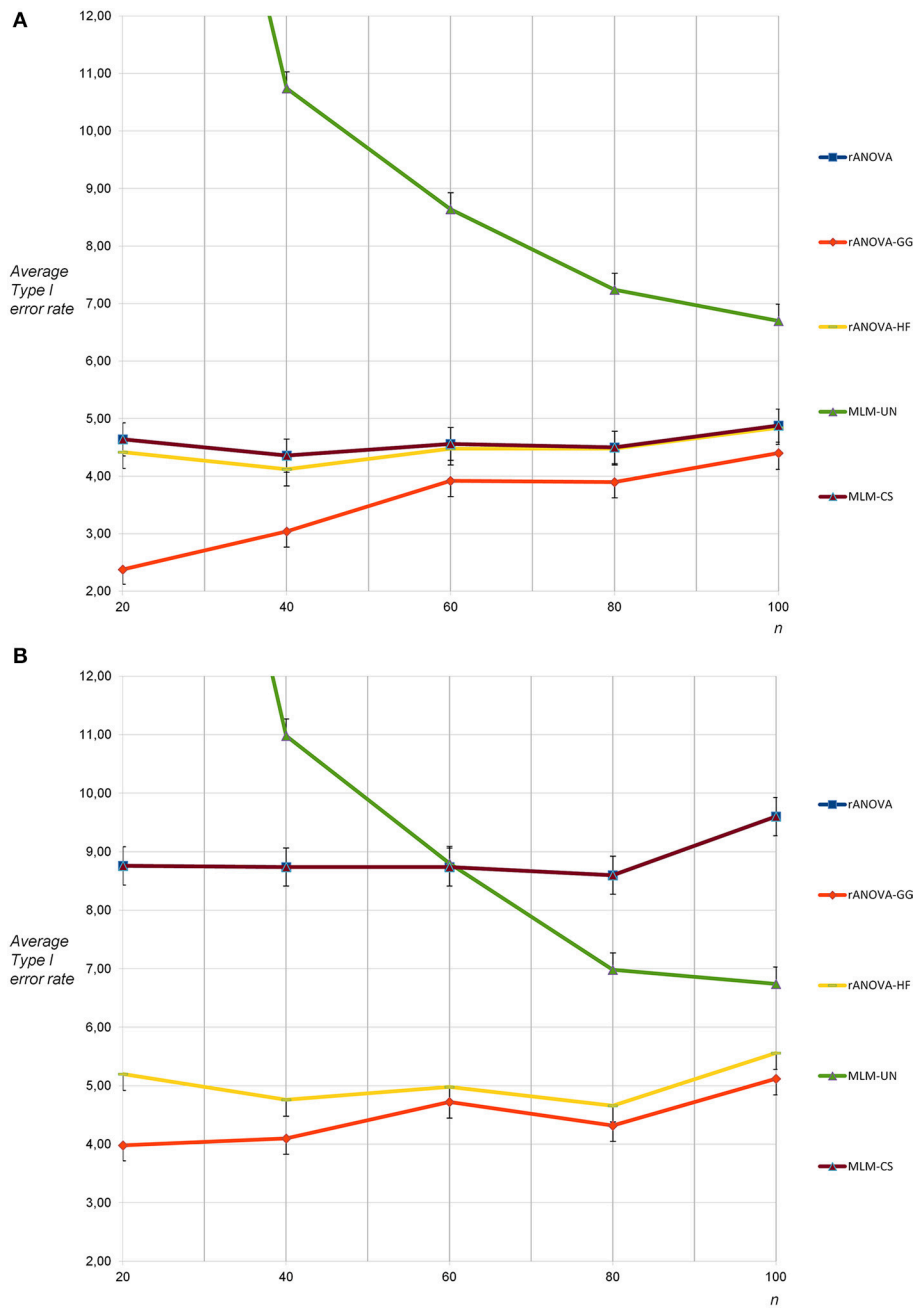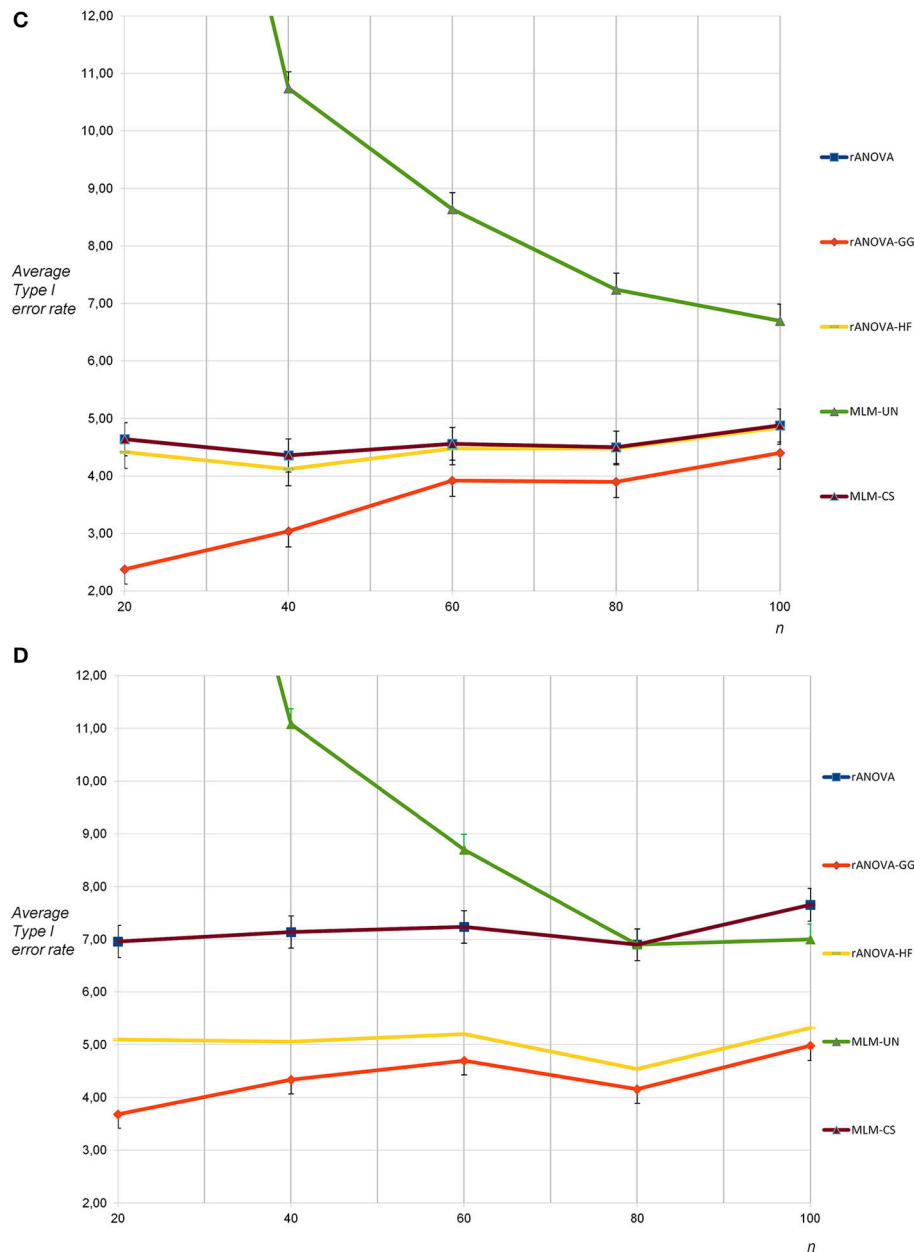

FIGURE 3 |

FIGURE 3 | (A) Average Type I error rates for 5,000 tests: no sphericity violation, nine measurement occasions, uncorrelated dependent variables (average rate for MLM-UN and $n = 20$: 22.72). **(B)** Average Type I error rates for 5,000 tests: sphericity violation, nine measurement occasions, uncorrelated and highly correlated dependent variables (average rate for MLM-UN and $n = 20$: 23.18). **(C)** Average Type I error rates for 5,000 tests: no sphericity violation, nine measurement occasions, moderately correlated dependent variables (average rate for MLM-UN and $n = 20$: 22.72). **(D)** Average Type I error rates for 5,000 tests: sphericity violation, nine measurement occasions, moderately correlated and highly correlated dependent variables (average rate for MLM-UN and $n = 20$: 23.38).

when there are nine or more measurement occasions and when MLM-UN is used because the sphericity assumption is violated in the data, sample sizes of at least 80 participants should be investigated.

The simulated data of Goedert et al. (2013) and Gueorguieva and Krystal (2004) were based on a violation of the sphericity assumption and on a substantially smaller number of measurement occasions. Therefore, the substantial Type I

error rates of MLM-UN in the condition without a violation of the sphericity assumption, nine measurement occasions and sample sizes of $n = 60$ and below could not be found in these simulation studies. Since the Type I error rates of MLM-CS were correct under this condition, this result indicates that MLM-UN should not be used as a form of standard procedure and that a specification of a known covariance structure in MLM might help to avoid progressive bias.

The results for MLM-CS and (uncorrected) rANOVA were so similar across all conditions of the simulation study that there was a total overlap of the respective lines in the figures. Accordingly, when the sphericity assumption was violated in the population data, MLM-CS had the same progressive bias as rANOVA. The progressive bias of MLM-CS and rANOVA increased slightly with the number of measurement occasions so that it was even larger than 0.075 ($=1.5\alpha$) for nine measurement occasions. Nevertheless, the progressive bias of MLM-CS and rANOVA was already substantial for three measurement occasions. Accordingly, MLM-CS and rANOVA cannot be recommended when the sphericity assumption is violated. It should be noted that MLM-UN resulted in more correct Type I error rates than MLM-CS and rANOVA when the sphericity assumption was violated and when sample size was $n = 80$ or larger. However, our results differ from Kowalchuk et al.'s (2004) finding that MLM-UN performs similarly to fitting the true covariance structure and, under certain conditions, shows even better Type I error control because MLM-CS had a more correct Type I error rate than MLM-UN when the sphericity assumption was not violated.

There was a conservative bias of rANOVA-GG, especially when the population data was conform to the sphericity assumption, when the number of measurement occasions was large, and when the sample size was small. Therefore, the use of rANOVA-GG cannot be recommended with $n = 20$ and nine measurement occasions when the sphericity assumption is not violated. rANOVA-HF had equal or more correct Type I error rates across all conditions of the simulation study than rANOVA-GG. Thus, the present simulation study supports the use of rANOVA-HF instead of rANOVA-GG.

An empirical example referred to the test of overall mean differences between items for the measurement of reward sensitivity. The sphericity assumption was violated in these data. rANOVA-GG, rANOVA-HF, and MLM-UN yielded similar results. The example illustrates method convergence with nine measurement occasions and more than 80 participants.

The present study was limited to the analysis of Type I error rates of within-group effects in data and models that do not contain any between-group effect. The aim of this restriction was to eliminate any interaction of between-group effects with within-group effects that might affect the Type I error rates of the within-group effects. Although this might be regarded as a limitation of the present study, the effects reported here have the merit of being independent from any between-group effect in the

data and in the models for data analysis. Further research may consider the investigation of between-group effects and especially the investigation of between-group × within-group interaction effects for up to nine measurement occasions.

Moreover, the statistical power of MLM and rANOVA should also be investigated for a large number of measurement occasions. There are, of course, several other issues, as, for example, the combined effect of the number of data levels and measurement occasions on Type I error rates and statistical power. The high flexibility of MLM as a method of multivariate analysis results in a specific responsibility of researchers in the application of this method. However, given a responsible use of MLM, the considerable flexibility is an important advantage of this method.

Two general recommendations follow from the results of the present simulation study with respect to the violation of the sphericity assumption: (1) Use rANOVA-HF, especially when sample sizes are small and when the number of measurement occasions is large. (2) Use MLM-UN when the sample size is at least $n = 80$ when there are six or more measurement occasions. Accordingly, when there are six or more measurement occasions and when there are reasons for performing MLM-UN instead of rANOVA-HF as, for example, when there is a substantial number of missing values or a more complex multi-level structure of the data, sample sizes of at least $n = 80$ should be investigated. Whether the sphericity assumption is violated or not, MLM-UN should not be used in combination with sample sizes of about $n = 60$ or smaller with nine or more measurement occasions unless a correction of the substantial progressive bias of this method is available. However, the proportionality of progressive bias of MLM-UN with the number of measurement occasions can be compensated by means of large sample sizes.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

SUPPLEMENTARY MATERIAL

REFERENCES

Arnau, J., Balluerka, N., Bono, R., and Gorostiaga, A. (2010). General linear mixed model for analysing longitudinal data in developmental research. *Percept. Mot. Skills* 110, 547–566. doi: 10.2466/pms.110.2.547-566

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Boisgontier, M. P., and Cheval, B. (2016). The anova to mixed model transition. *Neurosci. Biobehav. Rev.* 68, 1004–1005. doi: 10.1016/j.neubiorev.2016.05.034

Bradley, J. V. (1978). Robustness? *Br. J. Math. Stat. Psychol.* 31, 144–152.

Bryk, A. S., and Raudenbush, S. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage Publications.

Field, A. (1998). A bluffer's guide to sphericity. *Brit. Psychol. Soc. Math. Stat. Comput. Section Newsl.* 6, 13–22.

Goedert, K. M., Boston, R. C., and Barrett, A. M. (2013). Advancing the science of spatial neglect rehabilitation: an improved statistical approach with mixed linear modeling. *Front. Hum. Neurosci.* 7:211. doi: 10.3389/fnhum.2013.00211

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

Gueorguieva, R., and Krystal, J. H. (2004). Move over ANOVA. Progress in analyzing repeated-measures data and its reflection on papers published

in the Archives of General Psychiatry. *Arch. Gen. Psychiatry* 61, 310–317. doi: 10.1001/archpsyc.61.3.310

Hox, J. J. (2002). *Multilevel Analysis: Techniques and Applications.* Mahwah, NJ: Erlbaum.

Huynh, H., and Feldt, L. S. (1976). Estimation of the Box for degrees of freedom from sample data in randomised block and split-plot designs. *J. Educ. Stat.* 1, 69–82. doi: 10.2307/1164736

Knuth, D. E. (1981). *The Art of Computer Programming, 2nd Edn.*, Vol. 2. Reading, MA: Addison-Wesley.

Kowalchuk, R. A., Keselman, H. J., Algina, J., and Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted F tests. *Educ. Psychol. Meas.* 64, 224–242. doi: 10.1177/0013164403260196

Krishnamoorthy, K. (2006). *Handbook of Statistical Distributions with Applications.* New York, NY: Chapman & Hall/CRC. doi: 10.1201/9781420011371

Lucas, S. R. (2014). An inconvenient dataset: bias and inappropriate inference with the multilevel model. *Qual. Quant.* 48, 1619–1649. doi: 10.1007/s11135-013-9865-x

Maas, C. J. M., and Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology* 1, 86–92. doi: 10.1027/1614-2241.1.3.86

McLean, R. A., Sanders, W. L., and Stroup, W. W. (1991). A unified approach to mixed linear models. *Am. Stat.* 45, 54–64.

Murdoch, D. J., Tsai, Y.-L., and Adcock, J. (2008). P-values are random variables. *Am. Stat.* 62, 242–245. doi: 10.1198/000313008X332421

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd Edn.* Thousand Oaks, CA: Sage Publications.

Tabachnick, B. G., and Fidell, L. S. (2013). *Using Multivariate Statistics, 6th Edn.* Edinburgh Gate: Pearson Education.

West, B. T., Welch, K. B., and Galecki, A. T. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software.* Boca Raton, FL: Chapman & Hall/CRC.