Simplified estimation and testing in unbalanced repeated measures designs: Supplementary Material

1 Introduction

In this companion paper to the article 'Simplified estimation and testing in unbalanced repeated measures designs' (main text), we present additional results and information about the task switching experiment. In the first part, we describe a simulation study to compare the Wald tests discussed in the main text for main and interaction effects with F-tests from the standard repeated measures analysis of variance (ANOVA). In the second part, results for the real data set using the simple GEE-based approach are compared with results from F-tests adopting a standard repeated measures ANOVA. Finally, part three presents a more detailed description but also some additional results for the task switching experiment described in the main text.

2 Comparison of the simple GEE-based approach with repeated measures ANOVA: Simulation Study

In an additional simulation experiment, we compared the simple GEE approach proposed in the main text with results from a standard repeated measures ANOVA based on Type III sum of squares. Following the standard approach, a first step is to calculate individual cell means and then within-subject F-statistics. For justified valid inferences based on these univariate F-tests, however, the error covariance matrix needs to be circular — which is equivalent to sphericity of the covariance matrix that results from an orthonormal transformation of the error covariance matrix — and the errors are assumed to be normally distributed (e.g., Winer, Brown and Michels, 1991). If a F-statistic is significant, the assumption of sphericity is often tested by Mauchly's sphericity test (Mauchly, 1940) and, if this test is significant, a corrected p-value is interpreted. Two corrections are routinely used, the Greenhouse-Geisser (Greenhouse and Geisser, 1959) and the Huynh-

Feldt (Huynh and Feldt, 1976) correction, where the latter is less conservative than the former (Huynh and Feldt, 1976).

The simple estimator proposed in the main paper is independent from the error covariance structure. Since the Wald (or χ^2-) test and confidence intervals based on the t-distribution are justified assuming normality of the errors, we propose bootstrap versions for the Wald test and confidence intervals for linear combinations of parameters to avoid the normality assumption.

The simulation study considered here is similar to the one described in the main text. However, there are five differences: Firstly, to avoid long computing times we consider a model with only two factors (A and B) and their interaction (A*B), where A has three and B has five levels. Secondly, under the non-normality condition, all three error components were generated from an inverse Gaussian distribution with parameters $\mu = 1$ and $\lambda = 0.2$ and, unlike in the main text, signs were not reversed. Again, the realized skewness and excess kurtosis of the error terms $\sigma_k v_{i,k,t}$ over the simulations will be reported together with the simulation results. Thirdly, the variation of the true variances is larger: In the current simulation study we generated variances randomly from a uniform distribution with minimum 16 and maximum 127 (in the main text the minimum was one and the maximum was four). This implements a more extreme situation as compared to the simulations described in the main text. The fourth difference is due to the different restrictions imposed on the true effects in the main text (asymmetric restriction) as compared to the restrictions adopted in the R function (R Core Team, 2018) used to estimate the repeated measures model (Anova, from the `car` package, see Fox and Weisberg, 2011). In the latter case symmetric restrictions are adopted and thus we realized these in matrix $\mathbf{Z}$ as well. Finally, all true effects in $\boldsymbol{\theta}$ were set to zero.

As in the simulations reported in the main text all results are based on 500 simulated data sets for a realized low (low corr., $\rho_1 = .25$ and $\rho_2 = .3$) and high correlation condition (high corr., $\rho_1 = .5$ and $\rho_2 = .8$) and randomly varying $T_{i,k}$ (over i and k), but fixed over the simulations. The results under each condition are shown in Table 1 for $n = 33$ and in Table 2 for $n = 200$.

Table 1: Actual non-rejection rates of true 'no-effects'-null hypotheses based on the simple GEE estimator using the standard Wald test (W) and the bootstrap version (W^*), and based on repeated measures ANOVA F-tests (F) and, if Mauchly's sphericity test (M) is significant, based on corrected F-tests (F_{GG} for the Greenhouse-Geisser and F_{HF} for the Huynh-Feldt correction). Results are for $n = 33$, errors are functions of normally or inverse Gaussian distributed variables with skewness (skew) and excess kurtosis (exkurt), $\alpha = 0.05$.

	Effects	Non-rejection rates based on					
		W	W^*	F	M	F_{GG}	F_{HF}
		normally distributed error components					
low corr.	A	0.924	0.952	0.944	0.930	0.944	0.944
skew: -0.00	B	0.898	0.956	0.948	0.914	0.948	0.948
exkurt: 0.31	A*B	0.772	0.954	0.936	0.696	0.942	0.938
high corr.	A	0.920	0.954	0.940	0.934	0.940	0.940
skew: -0.01	B	0.910	0.954	0.950	0.932	0.952	0.952
exkurt: 0.26	A*B	0.778	0.948	0.952	0.796	0.956	0.954
		inverse Gaussian distributed error components					
low corr.	A	0.932	0.958	0.956	0.866	0.958	0.956
skew: 5.03	B	0.906	0.958	0.966	0.682	0.974	0.968
exkurt: 47.71	A*B	0.764	0.946	0.940	0.262	0.962	0.952
high corr.	A	0.936	0.964	0.950	0.694	0.960	0.958
skew: 3.37	B	0.896	0.954	0.960	0.342	0.976	0.972
exkurt: 18.84	A*B	0.728	0.964	0.952	0.030	0.986	0.980

The results in Table 1 imply that, as in the main text, whereas the rates of falsely rejecting the true null hypothesis by the standard Wald test systematically increases with an increasing number of parameters involved, the bootstrap version of the Wald statistic based on the simple GEE approach works well, i.e., the actual non-rejection rates are acceptable throughout whether the error components are normally distributed or follow an inverse Gaussian distribution. The same is true, however, for the (univariate) F statistic of the repeated measures model despite the fact that Mauchly's sphericity test implies violations of the sphericity assumption. For the normally distributed error components the inferences based on the Greenhouse-Geisser correction and on the Huynh-Feldt correction are almost the same and both seem to be valid. The general results are similar under the condition of error components following an inverse Gaussian distribution, although the rejection rates of the null hypothesis of sphericity are systematically higher and both, the Greenhouse-Geisser and Huynh-Feldt criteria seem to slightly over-correct for the interaction term A*B. Note that under the non-normal condition skewness and excess kurtosis are clearly distinct from zero.

Table 2: Actual non-rejection rates of true 'no-effects'-null hypotheses based on the simple GEE estimator using the standard Wald test (W) and the bootstrap version (W*), and based on repeated measures ANOVA F-tests (F) and, if Mauchly's sphericity test (M) is significant, based on corrected F-tests (F_{GG} for the Greenhouse-Geisser and F_{HF} for the Huynh-Feldt correction). Results are for $n = 200$, errors are functions of normally or inverse Gaussian distributed variables with skewness (skew) and excess kurtosis (exkurt), $\alpha = 0.05$.

		Non-rejection rates based on					
	Effects	W	W*	F	M	F_{GG}	F_{HF}
	normally distributed error components						
low corr.	A	0.958	0.964	0.968	0.772	0.968	0.968
skew: −0.00	B	0.932	0.938	0.952	0.694	0.952	0.952
exkurt: 0.32	A*B	0.938	0.954	0.956	0.000	0.962	0.962
high corr.	A	0.954	0.954	0.958	0.850	0.958	0.958
skew: −0.00	B	0.930	0.938	0.936	0.768	0.938	0.936
exkurt: 0.31	A*B	0.924	0.934	0.952	0.004	0.960	0.956
	inverse Gaussian distributed error components						
low corr.	A	0.944	0.946	0.954	0.704	0.954	0.954
skew: 5.00	B	0.946	0.952	0.964	0.424	0.968	0.966
exkurt: 46.04	A*B	0.924	0.946	0.942	0.000	0.950	0.944
high corr.	A	0.966	0.968	0.962	0.564	0.962	0.962
skew: 3.74	B	0.930	0.944	0.958	0.176	0.960	0.960
exkurt: 22.14	A*B	0.888	0.934	0.928	0.000	0.950	0.946

In the larger sample size condition ($n = 200$, see Table 2), the results are similar: The bootstrap version of the Wald test leads to acceptable non-rejection rates throughout, but also the repeated measures F-tests. Under the non-normal condition, the α-error seems to be slightly too high for the interaction term, but it can not be decided whether this is a systematic result or not. Mauchly's test indicates a violation of sphericity in even more cases as for the small sample condition but, again, even more so under the non-normal condition as compared to the normal condition.

The results presented in this section imply that for the conditions realized in this study, as already illustrated in the main text, the bootstrap version of the Wald test leads to acceptable coverage rates and thus seems to be insensitive to the number of parameters involved but also also shows some robustness if the errors do not follow the normal distribution. Interestingly, the (univariate) F statistic of standard repeated measures ANOVA seems to be remarkably robust with respect to violations of the circularity assumption of the error covariance matrix, but also against misspecification of the normality assumption. Note that only a very restricted set of possible simulations is realized in these simulations. It would be interesting to run more extensive and systematic simulations varying the true correlation matrix and realizing more extreme situations with respect to the variation of the true variances, the distributions (skewness and excess kurtosis) and allowing for individually varying error variances as well.

3 Comparison of simple GEE-based approach with repeated measures ANOVA: Application

In this section we compare results for the task switching experiment, described in the main text and in the next section in more detail, adopting the simple GEE-based approach and the standard repeated measures ANOVA approach. Table 3 presents the Wald statistics (Wald) with critical values based on the normal (crit_N) and the bootstrap (crit^*) distribution (see main text, $\alpha = 0.05$), the (univariate) F statistics of the repeated measures ANOVA (F) with corresponding p-values (P_F), Mauchly's test statistics of sphericity (M) and their p-values (P_M), the Greenhouse-Geisser and Huynh-Feldt criteria (GG and HF) and their corresponding p-values (P_{GG} and P_{HF}).

Note that the results presented in Table 3 are not directly comparable to the results in the main text because here, due to the function Anova (Fox and Weisberg, 2011), symmetric restrictions on parameters are adopted but in the main text, in accordance with the hypotheses of scientific interest, asymmetric restrictions are used. Further note that Mauchly's sphericity test makes only sense if a factor or interaction has more than two levels. Therefore, the corresponding cells in Table 3 are left blank.

According to the Wald test and regardless of whether we compare the Wald test statistic with the 0.95-percentile based on the assumption of normality or on the bootstrap distribution, the following factors and interactions would be interpreted as significant

Table 3: Tests of factors and interactions of the $(4 \times 2 \times 2 \times 2)$ unbalanced repeated measures model (1) based on $n = 33$, testing the null hypotheses of no effects: Wald test statistic (Wald) and critical values assuming normality (crit_N) or based on bootstrap percentiles (crit*), F statistic (F) and corresponding p-value P_F, Mauchly's sphericity criterion (M) and corresponding p-value (P_M), Greenhouse-Geisser criterion (GG) and corresponding p-value (P_{GG}), Huynh-Feldt criterion (HF) with corresponding p-value (P_{HF}), $\alpha = 0.05$.

Factor	Wald	crit_N	crit*	F	P_F	M	P_M	GG	P_{GG}	HF	P_{HF}
intercept	844.334	3.841	4.339	818.748	0.000						
CTI	6.540	7.815	9.229	3.345	0.022	0.688	0.043	0.803	0.032	0.874	0.028
Task	95.529	3.841	4.615	92.634	0.000						
CTI*Task	1.056	7.815	9.591	0.485	0.694	0.778	0.173	0.877	0.669	0.963	0.687
Seq	36.890	3.841	4.347	35.773	0.000						
CTI*Seq	28.631	7.815	8.874	10.860	0.000	0.925	0.791	0.948	0.000	1.050	0.000
Task*Seq	1.482	3.841	4.011	1.437	0.239						
CTI*Task*Seq	0.647	7.815	9.364	0.233	0.873	0.918	0.757	0.948	0.863	1.051	0.873
Aff	121.394	3.841	4.415	117.715	0.000						
CTI*Aff	7.312	7.815	9.652	1.968	0.124	0.712	0.065	0.850	0.135	0.930	0.129
Task*Aff	2.381	3.841	3.827	2.309	0.138						
Seq*Aff	6.370	3.841	3.956	6.177	0.018						
CTI*Task*Aff	0.175	7.815	10.040	0.052	0.984	0.844	0.390	0.901	0.978	0.992	0.984
CTI*Seq*Aff	12.227	7.815	9.059	3.390	0.021	0.839	0.372	0.902	0.025	0.993	0.021
Task*Seq*Aff	13.752	3.841	3.864	13.335	0.001						
CTI*Task*Seq*Aff	2.900	7.815	10.248	1.358	0.260	0.815	0.278	0.886	0.262	0.974	0.261

(excluding the intercept, $\alpha = 0.05$): Task, Seq, CTI*Seq, Aff, Seq*Aff, CTI*Seq*Aff and Task*Seq*Aff. All other factors and interactions would be interpreted as insignificant.

Considering the results from the standard repeated measures Anova, we would first conclude that Mauchly's sphericity test is significant only for CTI in which case we would arrive at the same conclusion regardless of whether inferences would be based on F, GG or HF. In all other cases sphericity needs either not to be tested because the involved factors have only two levels – note that if factors have only two levels, then for each factor and their interactions only one parameter is estimated – or the test is insignificant. Thus, only F and P_F need to be interpreted. Comparing inferences from F and P_F with those from the Wald test, imply with but one exception the same conclusions, i.e. significance of Task, Seq, CTI*Seq, Aff, Seq*Aff, CTI*Seq*Aff and Task*Seq*Aff. In all these cases, Mauchly's sphericity test is not significant ($\alpha = 0.05$), so we could interpret the standard F-tests. The exception is CTI, which would not be interpreted as significant based on the Wald test but deemed to be significant based on the F-test. In this case also Mauchly's sphericity test is significant ($\alpha = 0.05$) and thus, we would interpret the test based on the Huynh-Feldt correction, which leads to more liberal decisions than the Greenhouse-Geisser correction. However, in both cases we would classify CTI as significant.

4 Application: Additional Material from the Task Switching Paradigm

In the next two subsections we describe the task switching experiment in more detail (next section) and show the estimation results for each of the estimated parameters based on the simple GEE-approach in Section 5 and adopting asymmetric restrictions for the true effects as in the main text.

4.1 Description of the experiment

Participants. Twenty-four female and nine male students of the University of Hamburg participated in the experiment. They ranged in age from 17 to 44 years.

Apparatus and stimuli. The participants viewed the screen from a distance of about 60 cm. The stimuli were presented in white color on a dark gray background and occurred inside a permanent frame made up of four horizontally arranged rectangles, extending 2.2 cm vertically and 8.0 cm horizontally. On each trial, one of two different tasks was presented. In the Position Task, a white disc or one of the digits 1, 2, 3 or 4 occurred in one of the rectangles and the participants indicated its location. The digits 1 to 4 also served as stimuli in the Number Task. Either a single digit was displayed in one of the rectangles or all four rectangles were filled with instances of the same digit. The participants judged the numeral value of the displayed digit(s). There were thus two types of stimuli. The first type of stimuli afforded only the currently relevant task (a white disc afforded the Position Task but not the Number Task whereas a homogenous string of four digits afforded the Number Task but not the Position Task.) We refer to these stimuli as single-affordant stimuli. By contrast, a single digit, presented in one of the rectangles afforded both tasks. We refer to these stimuli as dual-affordant stimuli. The participants responded by pressing the keys C, V, N, and M of a standard QWERTZ keyboard. In the Position Task, each possible stimulus location was assigned to the spatially corresponding response key. In the Number Task, the values 1 to 4 were assigned to the response keys from left to right. The participants pressed the response keys with the index fingers and the middle fingers.

Procedure. Each trial started with the presentation of a cue that indicated the upcoming task. The cue for the position task was the word "Position", presented in red color slightly above the four rectangles. In addition the outline of the rectangles turned red. The cue for the Number Task was the word "Zahl" (German for "number"), presented in cyan. In addition, the outline of the rectangles turned cyan. The interval between the onset of the task cue and the presentation of the imperative stimulus (cue-target interval, CTI) was randomly drawn from the values 300 ms, 600 ms, 900 ms, and 1,200 ms. The stimulus was chosen from the set of single-affordant stimuli of the currently relevant task and the dual-affordant stimuli, with a probability of 50% each. Except for the constraint that (dual-affordant) stimuli associated with the same response in both tasks were never administered (i.e., the digit 1 in the leftmost rectangle or the digit 3 in the third rectangle from the left), the stimulus was chosen randomly from the set of possible stimuli. Task cue and imperative stimulus disappeared from the screen when the response was given. At the same time the outline of the rectangles turned white. If the response was correct, the cue of the subsequent trial occurred after 1,000 ms. If the response was incorrect, the word "Fehler" (German for "error") was displayed for 1,000 ms, and the cue of the subsequent trial occurred 1,000 ms later. The trial was then repeated with the same task and stimulus. These repetitions of incorrect trials were not counted as trials and response data were not included in the analyses. At the beginning of each session, the participant practiced the Position Task and the Number Task in single-task blocks of 99 trials each. They were then administered one block of 99 trials and three blocks of 100 trials each, in which the task was chosen randomly on each trial. Only data from these mixed blocks were subjected to the analyses. Between blocks, the participants were allowed to rest for some time. An experimental session lasted approximately 30 minutes.

5 Detailed Analysis Results

Table 4 presents individual results for each estimated parameter based on the simple GEE approach with asymmetric restrictions, i.e. defining a reference category for the levels of each factor, as described in the main text. Note that these estimation results must be expected to be different from those one would get from imposing symmetric restrictions on the parameters, i.e. considering for each factor the deviation of level effects from the mean over the level effects, as realized in Sections 2 and 3.

Table 4: Estimation results of the $(4 \times 2 \times 2 \times 2)$ repeated measures model from the main text (1) based on $n = 33$: Estimates $\hat{\theta}_k$ (estimate), $k = 1, \ldots, 32$, their standard errors (se), z-values (z-value), lower (ci.low) and upper (ci.up) confidence interval bounds and p-value (p-value), $\alpha = 0.05$.

variable	estimate	se	z-value	ci.low	ci.up	p-value
intercept	516.9905	14.7412	35.0712	488.0983	545.8826	0.0000
θ_{600}	-0.7912	8.1968	-0.0965	-16.8566	15.2742	0.9231
θ_{900}	-0.2514	10.1982	-0.0247	-20.2395	19.7367	0.9803
θ_{1200}	0.9784	8.9980	0.1087	-16.6573	18.6140	0.9134
θ_{dig}	85.4212	13.2684	6.4380	59.4157	111.4268	0.0000
$\theta_{600,\text{dig}}$	12.9229	18.1336	0.7127	-22.6183	48.4641	0.4761
$\theta_{900,\text{dig}}$	13.0714	16.9213	0.7725	-20.0938	46.2367	0.4398
$\theta_{1200,\text{dig}}$	16.7286	16.8188	0.9946	-16.2357	49.6928	0.3199
θ_{alt}	57.2044	12.3060	4.6485	33.0852	81.3236	0.0000
$\theta_{600,\text{alt}}$	-5.5401	16.8569	-0.3287	-38.5790	27.4988	0.7424
$\theta_{900,\text{alt}}$	-30.8124	13.7202	-2.2458	-57.7035	-3.9213	0.0247
$\theta_{1200,\text{alt}}$	-31.2154	14.7692	-2.1135	-60.1625	-2.2683	0.0346
$\theta_{\text{dig,alt}}$	31.3040	17.9190	1.7470	-3.8165	66.4246	0.0806
$\theta_{600,\text{dig,alt}}$	-36.7450	25.9746	-1.4147	-87.6542	14.1643	0.1572
$\theta_{900,\text{dig,alt}}$	-8.0969	26.2245	-0.3088	-59.4960	43.3021	0.7575
$\theta_{1200,\text{dig,alt}}$	-11.6610	18.4552	-0.6319	-47.8324	24.5105	0.5275
θ_{inc}	92.0476	19.6348	4.6880	53.5641	130.5312	0.0000
$\theta_{600,\text{inc}}$	53.8792	22.3863	2.4068	10.0028	97.7556	0.0161
$\theta_{900,\text{inc}}$	36.5326	24.9800	1.4625	-12.4273	85.4924	0.1436
$\theta_{1200,\text{inc}}$	22.0054	27.2561	0.8074	-31.4156	75.4263	0.4195
$\theta_{\text{dig,inc}}$	36.0928	26.6406	1.3548	-16.1218	88.3074	0.1755
$\theta_{600,\text{dig,inc}}$	-50.3233	35.5537	-1.4154	-120.0072	19.3606	0.1569
$\theta_{900,\text{dig,inc}}$	-31.5342	31.2919	-1.0077	-92.8653	29.7969	0.3136
$\theta_{1200,\text{dig,inc}}$	-28.9583	31.9993	-0.9050	-91.6757	33.7591	0.3655
$\theta_{\text{alt,inc}}$	138.2424	27.1561	5.0907	85.0173	191.4674	0.0000
$\theta_{600,\text{alt,inc}}$	-130.5041	37.6885	-3.4627	-204.3721	-56.6360	0.0005
$\theta_{900,\text{alt,inc}}$	-106.2384	36.9029	-2.8789	-178.5667	-33.9101	0.0040
$\theta_{1200,\text{alt,inc}}$	-99.4266	44.4252	-2.2381	-186.4984	-12.3549	0.0252
$\theta_{\text{dig,alt,inc}}$	-115.3622	40.1147	-2.8758	-193.9854	-36.7389	0.0040
$\theta_{600,\text{dig,alt,inc}}$	113.7548	67.0411	1.6968	-17.6433	245.1529	0.0897
$\theta_{900,\text{dig,alt,inc}}$	54.5695	53.6331	1.0175	-50.5495	159.6885	0.3089
$\theta_{1200,\text{dig,alt,inc}}$	60.7300	51.3829	1.1819	-39.9786	161.4387	0.2372

References

Fox, J. & Weisberg, S. (2011, 2nd ed). *An R Companion to Applied Regression.* Thousand Oaks CA: Sage. `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`

Greenhouse, S.W. & Geisser, S. (1959). On Methods in the Analysis of Profile Data. Psychometrika, 24(2), 95–112.

Huynh, H. & Feldt, L.S. (1976). Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split-Plot Designs. *Journal of Educational Statistics, 1(1),* 69–82.

Mauchly, J.W. (1940). Significance Test for Sphericity of a Normal n-Variate Distribution. *Annals of Mathematical Statistics, 11*, 204–209.

R Core Team (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org`.

Winer, B.J., Brown, D.R. & Michels, K.M. (1991, 3rd ed.). *Statistical Principles in Experimental Design.* McGraw-Hill: New York.