The Mixed Assessor Model and the multiplicative mixed model

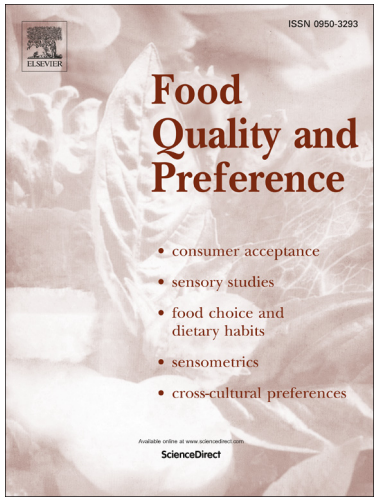
Pødenphant, Sofie; Truong, Minh H.; Kristensen, Kasper; Brockhoff, Per B.

Published in:
Food Quality and Preference

Link to article, DOI:
10.1016/j.foodqual.2018.11.006

Publication date:
2018

Document Version
Peer reviewed version



The Mixed Assessor Model and the multiplicative mixed model

Sofie Pødenphant, Minh H. Truong, Kasper Kristensen, Per B. Brockhoff

PII: S0950-3293(18)30650-5
DOI: https://doi.org/10.1016/j.foodqual.2018.11.006
Reference: FQAP 3599

To appear in: *Food Quality and Preference*

Please cite this article as: Pødenphant, S., Truong, M.H., Kristensen, K., Brockhoff, P.B., The Mixed Assessor Model and the multiplicative mixed model, *Food Quality and Preference* (2018), doi: https://doi.org/10.1016/j.foodqual.2018.11.006

The Mixed Assessor Model and the multiplicative mixed model

Sofie Pødenphant[a,*], Minh H. Truong[a], Kasper Kristensen[b], Per B. Brockhoff[a]

[a] *DTU Compute, Section for Statistics and Data Analysis, Technical University of Denmark, Richard Petersens Plads, Building 324, DK-2800 Kongens Lyngby, Denmark*
[b] *DTU Aqua, Section for Marine Living Resources, Technical University of Denmark, Kemitorvet, Building 202, DK-2800 Kongens Lyngby, Denmark.*

Abstract

A novel possibility for easy and open source based analysis of sensory profile data by a formal multiplicative mixed model (mumm) with fixed product effects and random assessor effects is presented by means of the generic statistical R-package *mumm*. The package is using likelihood principles and is utilizing newer developments within Automatic Differentiation by means of the Template Model Builder R-package. We compare such formal likelihood based analysis with the Mixed Assessor Model (MAM) analysis, where MAM is a linear approximation of the multiplicative mixed model. We use real sensory data as examples together with simulated data. We found that the formal mumm approach for hypothesis testing more resembles the MAM than the standard 2-way mixed model, and that both the mumm approach and the MAM give a higher power to detect product differences than the 2-way mixed model, when a "scaling effect" is present. We also validated that the novel contrast confidence limit method suggested previously for the MAM performs well and in line with the formal likelihood based confidence intervals of the mumm. Finally, the likelihood based mumm approach suggests that the more proper test for product difference would be a test that has a "joint product and scaling effect" interpretation.

*Corresponding author
Email address: sofp@dtu.dk (Sofie Pødenphant)

Keywords: Sensory profile data, Analysis of variance, Multiplicative mixed model, Scaling differences, Disagreement, Template Model Builder

1. Introduction

Sensory profile data, where I assessors scored J products in K replications, is frequently analysed by a 2-way mixed analysis of variance (ANOVA) corresponding to the following model

$$\begin{aligned} Y_{ijk} &= \mu + a_i + \nu_j + g_{ij} + \epsilon_{ijk} \\ a_i &\sim N(0, \sigma^2_{PAN}), g_{ij} \sim N(0, \sigma^2_G), \epsilon_{ijk} \sim N(0, \sigma^2), \end{aligned} \tag{1}$$

where a_i is the random assessor main effect, $i = 1, ..., I$, ν_j is the product main effect, $j = 1, ..., J$, g_{ij} is the random assessor-by-product interaction and ϵ_{ijk}, $k = 1, ..., K$, is the random residual error.

However, the assessor-by-product interaction will often not only consist of real deviations in perception of product differences (disagreement effect), but also of scale range differences between assessors (scale effect). Scale range differences appear when some assessors use a larger part of the scale than others, when scoring the products. In Brockhoff et al. (2015) a meta study of 8619 attributes from 369 profile data sets showed that such scaling heterogeneity was significantly present in 45% of all the attributes. Thus, it will not be valid in general to assume that scale range differences are not present. To account for the scale effect, Brockhoff et al. (2015) uses the multiplicative model approach suggested in Brockhoff & Skovgaard (1994) and combines it with the general mixed model approach. As a result, the Mixed Assessor Model (MAM) was introduced:

$$\begin{aligned} Y_{ijk} &= \mu + a_i + \nu_j + \beta_i x_j + d_{ij} + \epsilon_{ijk} \\ a_i &\sim N(0, \sigma^2_{PAN}), d_{ij} \sim N(0, \sigma^2_D), \epsilon_{ijk} \sim N(0, \sigma^2), \end{aligned} \tag{2}$$

where $x_j = \overline{y}_{.j.} - \overline{y}_{...}$ are the centered product averages inserted as a covariate, implying that the βs are the individual scaling slopes; the bigger the scale

range, the larger the slope. The purpose of including the term $\beta_i x_j$ is to model the assessors' individual ranges of scale use, such that the interaction, d_{ij}, captures the disagreement effect and not the scale effect. Consequently, the scaling heterogeneity between the assessors is removed from the assessor-by-product interaction, when used for hypothesis testing of product differences. Since the x_js are calculated directly from data, MAM is a linear mixed model and should be seen as an approximation of the following more properly specified mixed model (Brockhoff et al., 2015)

$$
\begin{aligned}
&Y_{ijk} = \mu + a_i + \nu_j + b_i\nu_j + d_{ij} + \epsilon_{ijk} \qquad (3)\\
&(a_i, b_i) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma^2_{PAN} & \rho\,\sigma_{PAN}\,\sigma_{SCALE} \\ \rho\,\sigma_{PAN}\,\sigma_{SCALE} & \sigma^2_{SCALE} \end{bmatrix}\right),\\
&d_{ij} \sim N(0, \sigma^2_D), \epsilon_{ijk} \sim N(0, \sigma^2),
\end{aligned}
$$

where the covariate is the true product effects, ν_j, and where the scaling slopes, b_i, are modelled as random effects. It is worth noting that in Brockhoff et al. (2015), the model is written without the correlation between a_i and b_i. But the model is in that paper only used to express product contrast variances and these contrasts are not affected by this correlation, so everything in Brockhoff et al (2015) is unchanged based on the model including the correlation stated here. However, since it impacts the optimization of the likelihood function for the model and the resulting parameters, we allow for a correlation between them. Due to the multiplicative term, $b_i\nu_j$, this model is a so-called multiplicative mixed model and does not belong to the class of linear mixed models. The model is not straightforward to estimate, since the unknown product values, ν_j, enter both the expectation and the variance structure. Therefore, the theory and the computations become simplified when using the MAM.

In Smith et al. (2003) a multiplicative mixed model is also applied for the analysis of sensory profile data, but since they use the model to examine assessor performances, they assume assessor effects to be fixed and product effects random. Under that assumption, the model has a factor analytic covariance

structure, which simplifies the estimation. The assumption of random product effects and fixed assessor effects is, however, not reasonable in our case, since our main focus is on the comparison of specific products, and not on the particular assessors in the panel.

Even though the MAM is an approximation, Brockhoff et al. (2015) justifies that the model produces valid hypothesis tests for overall product differences and also for post-hoc product difference testing. They further show that MAM increases the power to detect product differences, compared to a standard 2-way mixed model. The MAM is also used in Peltier et al. (2014) for monitoring assessor and panel performance, based partly on ideas from Brockhoff (2003). In Brockhoff & Belmonte (2018) a full overview of the use of the MAM is given, together with a demonstration of how to fit the model by the R-package *SensMixed*, (Kuznetsova et al., 2016b; R Core Team, 2017).

However, the MAM will, in general, fail to produce valid post-hoc product difference confidence intervals. Therefore Brockhoff et al. (2015) suggests a novel procedure to obtain appropriate product difference confidence intervals.

In this paper, we compare the hypothesis tests for overall product differences when using the MAM, the 2-way mixed model and the multiplicative mixed model, where the latter is fitted by the newly developed R-package *mumm* (Pødenphant & Brockhoff, 2016) by optimization of the likelihood function (Section 2). Further, we investigate whether the suggested procedure in Brockhoff et al. (2015) actually does produce appropriate product difference confidence intervals. We will do this by using the procedure to estimate confidence intervals for simulated data sets and, thereafter, calculating coverage probabilities for the estimated confidence intervals. Additionally, we will use R-package *mumm* to find the profile likelihood based confidence intervals for the product differences for the multiplicative mixed model, and calculate the resulting coverage probabilities. Subsequently we will compare the performance of the full likelihood approach and the novel method suggested by Brockhoff et al. (2015) (Section 3). In Section 4, we propose a new test for product difference for the MAM.

Finally, Section 5 includes a summary, a discussion about computation time and some final remarks.

2. Power to detect product differences

The hypothesis tests for overall product differences were conducted by F-tests for model (1) and (2) and by likelihood ratio tests for model (3). The product difference F-test based on the MAM differs from the F-test based on the 2-way mixed model by having $MS_{Disagreement}$ instead of $MS_{Interaction}$ as denominator in the F-statistic. This is the reason for the increased power to detect product differences (Brockhoff et al., 2015). The F-test based on the MAM was carried out by the use of R-package *SensMixed* (Kuznetsova et al., 2016b). The likelihood ratio test for product differences in (3) was performed by testing the reduced (null) model with no product effect

$$\begin{aligned} &Y_{ijk} = \mu + a_i + d_{ij} + \epsilon_{ijk} \\ &a_i \sim N(0, \sigma^2_{PAN}), d_{ij} \sim N(0, \sigma^2_D), \epsilon_{ijk} \sim N(0, \sigma^2), \end{aligned} \tag{4}$$

against the full model. The full model and its likelihood were found by using the R-package *mumm* (Pødenphant & Brockhoff, 2016), which is a wrapper of the Template Model Builder R-package (*TMB*) (Kristensen et al., 2016). *TMB* enables fast optimization of the Laplace approximation of the marginal log-likelihood function for the multiplicative mixed model. The "Laplace approximation" is a standard procedure within likelihood theory, often used for nonlinear mixed models as a way to approximate the complicated likelihood function, which otherwise is a complex multi-integral expression (Wolfinger, 1993; Vonesh, 1996). Fast optimization of the Laplace approximation is made possible with *TMB* through the use of Automatic Differentiation (AD). With this technique, *TMB* obtains the gradient, the Hessian, and the third order derivatives of the joint log-likelihood function with respect to the random effect coefficients, in a very time-efficient manner. These values are used to construct the Laplace approximation and its gradient, which are given as inputs to a

standard minimizer in R to finally perform the maximization of the likelihood function.

It is assumed that the test statistic from the likelihood ratio test follows a chi-squared distribution under the null hypothesis of the products being equal, but the number of degrees of freedom in the test is not obvious, due to the multiplicative term. When hypothesizing that the products are all equal, a $J-1$ degrees of freedom hypothesis, then the potential scaling differences also vanish from the model: When the products are all the same, there is no way that the assessors can range them differently. This is known in statistical likelihood theory as the problem of "nuisance parameters only present under the alternative". This kind of challenge has received attention over the last decades, see, e.g., Davies (1977); Ritz & Skovgaard (2005). The practical challenge is that one does not know the exact nor approximate distribution of the likelihood ratio test statistic. In sensory applications a similar situation and challenge is seen for the joint test of product difference in the corrected beta-binomial analysis of replicated difference testing data, cf. Brockhoff & Linander (2017). In other words, we do not know the best possible choice of number of degrees of freedom for the test.

However, we do know that, with a significance level at 5%, the false discovery rate should also be 5%. In the likelihood ratio test for product difference, we have therefore used the number of degrees of freedom that, in a simulation study under the null hypothesis, fulfills this criteria. In the simulation study we used 1000 data sets, simulated from the null model (4), where the parameters are set equal to the values we get from fitting (4) to the example data set in question.

2.1. Data example

We have analyzed the *TVbo* data set from the R-package *lmerTest* (Kuznetsova et al., 2016a, 2017). The data stems from sensory evaluations of Bang & Olufsen (BO) televisions, which were characterized by two design factors, *Picture* and *TVset*, with 4 and 3 levels, respectively. To compare the power of the methods

in an illustrative way, a data set in which the product effect is on the boundary of significance is preferred. In this example, we have therefore allowed ourselves to analyze a subset of the data to lower the significance of the product effect. Thus, we have discarded all the observations for the two last levels of *Picture*, such that this factor only has 2 levels. Afterwards, the two design factors were crossed, yielding 6 products in total. The products were evaluated by 8 assessors in 2 replications, yielding 16 observations per product, and the assessment was based on 15 attributes, of which we chose *Dim glass effect* as the response variable for this example. The *Dim glass effect* is big if parts of the picture seems dim/dull and the scale goes from none to a lot.

Table 1 shows the ANOVA table for this data set and Table 2 shows the result of the hypothesis tests for overall product differences. The approximate number of degrees of freedom in the likelihood ratio test is 9, which was found from the method described previously. In Table 1, we see that a significant assessor-by-product interaction is present, of which a big part is explained by the scaling effect, whereas the disagreement effect is non-significant. Observing the p-values for the product effect in Table 2, we see that significance at a 5% level was obtained for all the methods. However, we failed to find a significant product effect at a 1% level, when using the standard 2-way mixed model. To see if the results from this example illustrate a general tendency, we have conducted a simulation study, to be presented in the next section.

	SS	MS	DF	F	p-value
Assessor	92.52	13.22	7	5.38	0.0003
Product	32.66	6.53	5		
Interaction:	86.00	2.46	35	3.76	$< .0001$
Scaling	69.68	9.95	7	17.08	$< .0001$
Disagreement	16.32	0.58	28	0.89	0.6205
Error	31.37	0.65	48		

Table 1: ANOVA table for the *TVbo* data set with *Dim glass effect* as the response variable.

Model	Test	DF_1	DF_2	F	χ^2	p-value
2-way (1)	F-test	5	35	2.66		$3.87 \cdot 10^{-2}$
MAM (2)	F-test	5	28	11.21		$5.32 \cdot 10^{-6}$
mumm (3)	LRT	9			59.93	$1.38 \cdot 10^{-9}$

Table 2: Hypothesis tests for overall product differences with *Dim glass effect* as the response variable.

2.2. Simulation study

For the comparison of the hypothesis tests for overall product difference, 1000 data sets have been simulated from the multiplicative mixed model (3). Each data set contains scores given by 8 assessors to 6 products in 2 replications. The product effect parameters and the variance components were set equal to the parameter estimates obtained from fitting model (3) by the *mumm* R-package to the *TVbo* data set with *Dim glass effect* as the response variable (without the two last levels of *Picture*). The parameter estimates are given in Table 3.

$\mu + \nu_1$	$\mu + \nu_2$	$\mu + \nu_3$	$\mu + \nu_4$	$\mu + \nu_5$	$\mu + \nu_6$
2.2324	2.9687	3.5215	2.0347	2.0388	2.0665

σ	σ_{PAN}	σ_{SCALE}	σ_D	ρ
0.7299	0.9588	1.5193	$3.0291 \cdot 10^{-05}$	0.6924

Table 3: The product effect parameters and the variance components used to simulate of data for comparison of the hypothesis tests.

In this study, we found that in 962 cases out of 1000, the p-value for the MAM was smaller than for the 2-way mixed model. This number was a bit larger for the multiplicative mixed model, with 991 p-values being smaller than for the 2-way model. When looking at the overall product difference significance, with a significance level at 0.05, the MAM finds a significant product

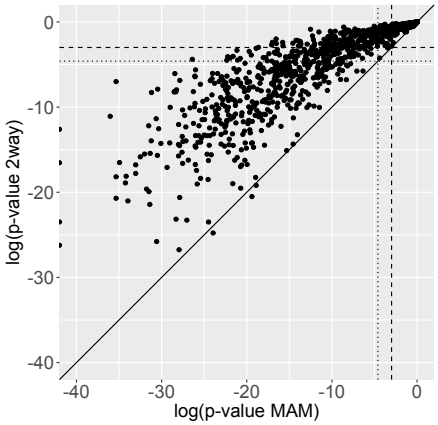


Figure 1: The logarithm of the p-values for the 2-way model plotted against the logarithm of the p-values for the MAM. The dotted and the dashed lines mark p-value = 0.01 and p-value = 0.05, respectively. For the dots on the boundary of the axis the computed value of log(p-value) is $-\infty$.

difference in 78.0% of the attributes, whereas the multiplicative mixed model finds a difference in 99.2% of the cases. The 2-way model finds the fewest significant differences, with a percentage of 55.8%. The difference between the methods becomes bigger if the significance level is lowered to 0.01. In this case, a significant difference is found for the MAM, the multiplicative mixed model and the 2-way model in 72.2%, 98.4% and 43.8% of the cases, respectively. This reflects the increased power obtained from using the MAM instead of the common 2-way mixed model, as described in Brockhoff et al. (2015). It further shows that using the multiplicative mixed model we obtain the greatest power to detect product differences.

To get a better understanding of the behavior of the p-values, Figure 1, 2 and 3 show scatter plots of the logarithm of the p-values. The log-transformation was chosen to "spread out" the values close to zero, since the p-values on to the border of significance are of most interest. The p-values for the 2-way model is plotted against the p-values for the MAM in Figure 1, which clearly illustrates that the p-value for the MAM in general is lower than for the 2-way

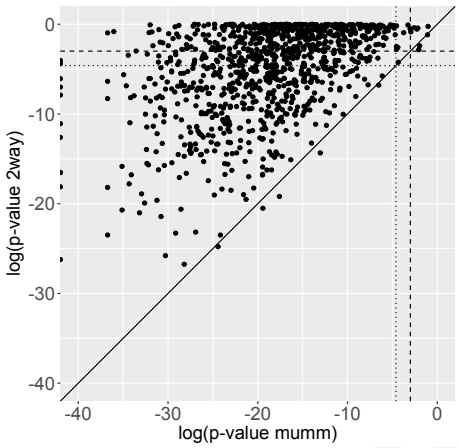


Figure 2: The logarithm of the p-values for the multiplicative mixed model plotted against the logarithm of the p-values for the 2-way model. The dotted and the dashed lines mark p-value = 0.01 and p-value = 0.05, respectively. For the dots on the boundary of the axis the computed value of log(p-value) is $-\infty$.

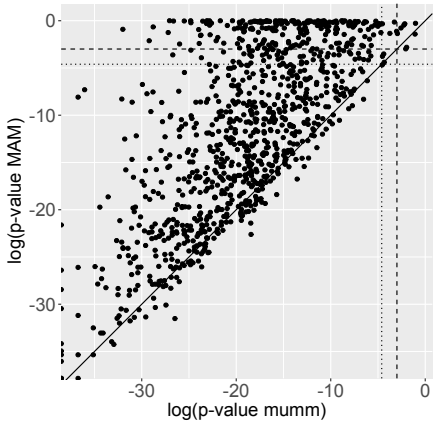


Figure 3: The logarithm of the p-values for the multiplicative mixed model plotted against the logarithm of the p-values for the MAM. The dotted and the dashed lines mark p-value = 0.01 and p-value = 0.05, respectively. For the dots on the boundary of the axis the computed value of log(p-value) is $-\infty$.

model. It is further observed that there are quite a lot of situations, where the p-value is above 0.05 for the 2-way model but below 0.01 for the MAM. Figure 2 shows the p-values for the 2-way model plotted against the p-values for the multiplicative mixed model. Here we also see a lot of situations, where the p-value is above 0.05 for the 2-way model but below 0.01 for the multiplicative mixed model. In figure 3, which shows the p-values for the MAM plotted against the p-values for the multiplicative mixed model, we see that the p-values for the multiplicative mixed model more resembles the p-values for the MAM than for the 2-way model. However, a similar pattern as before is seen, with a lot of situations, where the p-value is above 0.05 for the MAM but below 0.01 for the multiplicative mixed model. This indicates that using the multiplicative mixed model, we are able to capture some information about product differences that the two other models fail to. We elaborate more on this in Section 4.

2.3. Different data scenarios

In the previous data example, the scaling effect was present while the disagreement effect was non-significant. In this section, we want to investigate how generic our findings are, regarding the power to detect product differences, by considering two other types of data scenarios. In the first alternative example, we have analyzed a data set, where both the scaling effect and the disagreement effect are present (at a 5% level). The data also stems from the *TVbo* data set, but this time we choose the attribute *Colourbalance* as our response variable, and we have discarded all the observations for the first and the fourth level of *Picture*. In the last example, we have considered a data set, where the disagreement effect is present, while the scaling effect is non-significant. Yet again, the data stems from the *TVbo* data set, where we choose *Depth* as the response variable and discarded the observations for the second and third level of *Picture*.

Table 4 shows the p-values for the interaction effects and the overall product effect for the three data scenarios. When both the scaling effect and the disagreement effect are present, the p-value for an overall product effect is lower

for the MAM and for the multiplicative mixed model than for the 2-way model, with the MAM resulting in the lowest p-value. On the other hand, when the scaling effect is non-significant the standard 2-way mixed model gives the lowest p-value, when testing for an overall product effect. We have conducted

Data	ANOVA (interaction)			Product effect		
	Inter.	Scaling	Disagr.	2-way	MAM	mumm
1	$< .0001$	$< .0001$	0.6205	$3.87 \cdot 10^{-2}$	$5.32 \cdot 10^{-6}$	$1.38 \cdot 10^{-9}$
2	0.0016	0.0145	0.0413	0.0537	0.0145	0.0420
3	0.0049	0.5665	0.0051	$6.96 \cdot 10^{-3}$	$1.04 \cdot 10^{-2}$	$4.11 \cdot 10^{-2}$

Table 4: The p-values for the interaction effects and for the overall product effect in the three data sets. In data set 1, *Dim glass effect* is the response variable, in data set 2, *Colourbalance* is the response, and in data set 3, *Depth* is the response.

			Power					
Data	Comparison		$\alpha = 0.05$			$\alpha = 0.01$		
	1	2	2-way	MAM	mumm	2-way	MAM	mumm
1	96.2	99.1	55.8	78.0	99.2	43.8	72.2	98.4
2	82.5	63.9	66.7	75.2	77.8	47.5	59.3	60.1
3	26.9	7.2	93.4	92.6	83.2	79.1	75.5	61.2

Table 5: The results from the simulation studies. Comparison 1 states the percentage of times the p-value for the MAM is lower than the p-value for the 2-way model. Comparison 2 stats the percentage of times the p-value for the mumm is lower than for the 2-way model. The rest of the results are the percentage of times the models find a significant product effect (the power).

a simulation study for the two alternative data scenarios, similar to the one conducted for the first data set. The results are shown in Table 5. When the interaction consists of a significant scaling effect and a significant disagreement effect, the MAM and the multiplicative mixed model have a very similar power to detect product differences, whereas the 2-way model finds a significant effect less often. However, the difference between the methods are smaller in this sce-

nario, than in the scenario without a significant disagreement effect. When the scaling effect, on the other hand, is not significant, the 2-way mixed model has the largest power to detect product differences. This is anyhow not surprising, since the MAM and the multiplicative mixed model waste degrees of freedom on estimating a non-existing scaling effect.

3. Confidence intervals for product differences

The procedure suggested by Brockhoff et al. (2015) for obtaining product difference confidence intervals based on (2), while taking (3) into consideration, is implemented in the R-package *SensMixed*. The exact command used to achieve the estimated confidence intervals is shown in Appendix A. This novel procedure gives non-symmetrical confidence intervals, because the scaling variance is taken into account. This contradicts the confidence intervals one would obtain from running the MAM in standard statistical software for linear models, i.e computing the confidence intervals in the following way (Brockhoff et al., 2015):

$$\overline{y}_{.j.} - \overline{y}_{.j'.} \pm t_{0.975}(f)\sqrt{\frac{2MS_{Disagreement}}{IK}},$$

where $t_{0.975}(f)$ is the 0.975 quantile of the Student's t-distribution with degrees of freedom $f = DF_{Disagreement} = (I-1)(J-2)$.
It is important to note that this method is not recommended, since the scaling variance is falsely ignored. This method of computing confidence intervals will from now on be referred to as method "MAM$_{naive}$", whereas method "MAM" will denote the novel procedure suggested by Brockhoff et al. (2015), which takes the scaling variance into account.

The profile likelihood estimated confidence intervals based on (3) were found by R-package *mumm*. The *mumm* function calls, for obtaining the estimated confidence intervals, are shown in Appendix A. Let this method be referred to as method "mumm".

Note that none of the confidence intervals are corrected for multiple comparisons. Such adjustments could be implemented in the methods, but this is beyond the scope of this paper.

3.1. Data example

We consider again the *TVbo* data set, with the attribute *Dim glass effect* as the response variable, but this time we keep all of the observations, meaning that we now analyze 12 products.

Table 6 shows the ANOVA table for this data set and Table 7 shows the p-values for the product effect. We see that all the effects are highly significant, except for the disagreement effect, which is non-significant.

	SS	MS	DF	F	p-value
Assessor	277.11	39.59	7	9.05	$< .0001$
Product	295.76	26.89	11		
Interaction:	337.00	4.38	77	1.96	0.0009
Scaling	215.09	30.73	7	17.64	$< .0001$
Disagreement	121.90	1.74	70	0.78	0.8656
Error	214.89	2.24	96		

Table 6: ANOVA table for the *TVbo* data set with *Dim glass effect* as the response variable.

Model	Test	DF_1	DF_2	F	χ^2	p-value
2-way (1)	F-test	11	77	6.14		$3.87 \cdot 10^{-07}$
MAM (2)	F-test	11	70	15.44		$1.02 \cdot 10^{-14}$
mumm (3)	LRT	17			128.16	$< 10^{-14}$

Table 7: Hypothesis tests for overall product differences with *Dim glass effect* as the response variable.

It is clear that the product effect is significant no matter which model is used. We also see that the disagreement effect is non-significant, meaning that

the scaling effect alone explains the assessor-by-product interaction.

We have estimated the product contrasts and their corresponding 95%-confidence intervals by method "MAM", method "MAM$_{naive}$", method "mumm", and by using the standard 2-way mixed model. To illustrate the difference between these four methods, the contrast estimates and the confidence intervals for the three smallest and largest contrasts, ordered according to the estimates obtained from fitting model (3), are plotted in Figure 4. It is clear that the

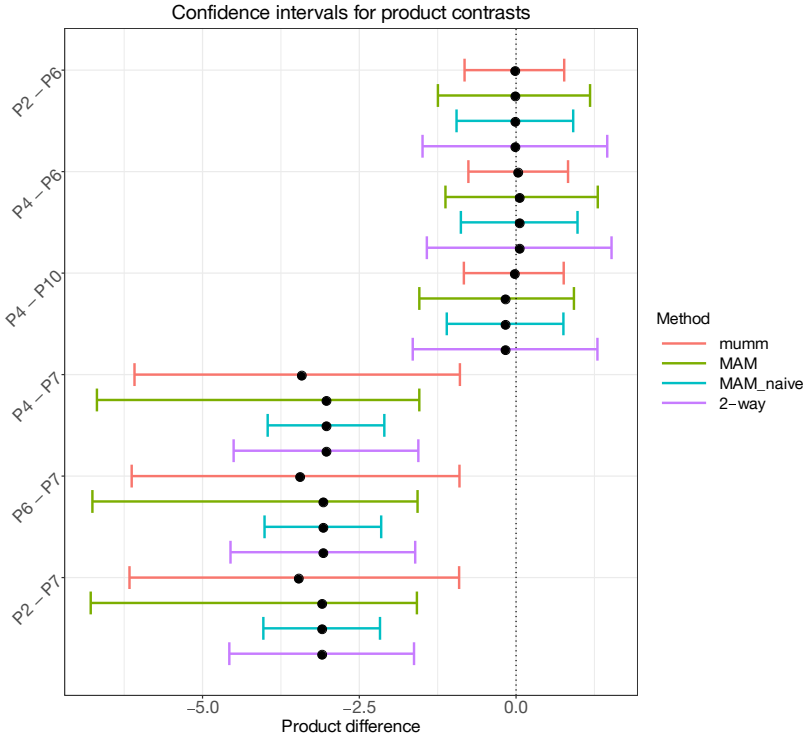


Figure 4: The estimated product contrasts (black dots) and the corresponding 95%-confidence intervals. Method "MAM" refers to the procedure suggested in Brockhoff et al. (2015), method "MAM$_{naive}$" refers to using MAM directly, and method "mumm" refers to the use of R-package *mumm* for finding the profile likelihood based confidence intervals. Only the confidence intervals for the three smallest and the three largest contrasts are plotted.

estimated confidence intervals from method "MAM" and method "mumm" are asymmetric as expected, being wider "away from zero" than "towards zero"

(Brockhoff et al., 2015). Moreover, the intervals obtained by method "MAM" are wider than the intervals obtained from method "mumm", for the small contrasts. The confidence intervals from method "MAM$_{naive}$" and the 2-way model are symmetric and have a constant width for all of the contrasts, which makes them notably shorter than the intervals from the two other methods, for large contrasts, and wider for small contrasts. For small contrasts, the intervals for "MAM$_{naive}$" look reasonable, whereas the intervals for the 2-way model appear to be too wide compared to the others. However, from this plot we cannot say much about which intervals are more correct. Therefore, we have conducted a simulation study to compute the coverage probabilities for the confidence intervals, which will be described in the following section.

3.2. Simulation study

To compute coverage probabilities for the estimated confidence intervals, 1000 data sets were simulated from the multiplicative mixed model (3). Each data set contains scores given by 8 assessors to 12 products in 2 replications. The product effect parameters and the variance components were set equal to the parameter estimates obtained from fitting model (3) to the *TVbo* data set with *Dim glass effect* as the response variable. The estimated model parameters used in the simulation study are shown in Table 8. We have estimated the

$\mu+\nu_1$	$\mu+\nu_2$	$\mu+\nu_3$	$\mu+\nu_4$	$\mu+\nu_5$	$\mu+\nu_6$
2.2259	1.9794	3.8116	2.0273	3.1773	2.0004

$\mu+\nu_7$	$\mu+\nu_8$	$\mu+\nu_9$	$\mu+\nu_{10}$	$\mu+\nu_{11}$	$\mu+\nu_{12}$
5.4519	2.4355	3.7462	2.0544	5.1984	2.0916

σ	σ_{PAN}	σ_{SCALE}	σ_D	ρ
1.3366	1.1700	0.9117	$6.62 \cdot 10^{-5}$	0.8025

Table 8: The model parameters used in the simulation of data sets.

product difference confidence intervals for all of the 1000 data sets by the four methods. Hereafter, the coverage probabilities of the confidence intervals have been calculated from the "true" known product differences.

Table 9 shows the coverage probabilities of the estimated confidence intervals, found from the simulation study. As seen, the coverage probabilities of the confidence intervals produced by method "MAM" look very reasonable and are very similar to the coverage probabilities for method "mumm". On the other hand, the coverage probabilities for method "MAM_{naive}" and for the 2-way model are clearly to high for small contrasts and they are unreasonably below 0.95 for large contrasts.

	Contrast	mumm	MAM	MAM_{naive}	2-way
P2-P6	-0.0210	0.9580	0.9670	0.9780	0.9940
P4-P6	0.0269	0.9600	0.9660	0.9830	0.9970
P4-P10	-0.0271	0.9430	0.9710	0.9830	0.9930
P10-P12	-0.0371	0.9550	0.9690	0.9790	0.9950
P2-P4	-0.0479	0.9530	0.9600	0.9760	0.9960
⋮	⋮	⋮	⋮		
P7-P12	3.3604	0.9350	0.9310	0.6940	0.7920
P7-P10	3.3975	0.9390	0.9270	0.7170	0.7990
P4-P7	-3.4246	0.9340	0.9310	0.6950	0.8000
P6-P7	-3.4515	0.9410	0.9310	0.6830	0.7780
P2-P7	-3.4725	0.9350	0.9380	0.6950	0.7850

Table 9: The coverage probablities.

3.3. Different data scenarios

To investigate how the results generalize to other data scenarios, we have estimated the product contrast confidence intervals for the *TVbo* data set with first *Colourbalance* and then *Depth* as the response variable. All the observations was included, yielding 12 products in total. As in Section 2.3, both

the scaling effect and the disagreement effect are present in the assessor-by-product interaction when *Colourbalance* is the response variable, whereas only the disagreement effect is present when *Depth* is chosen as the response variable. Figure 5 show the estimated confidence intervals for the first-mentioned example. We see that the difference between the methods is smaller now, compared to in Figure 4. The corresponding coverage probabilities are shown in Table 10, where we see that the two-way model and "MAM_{naive}" are still performing poorly - especially for large contrasts. We further see that method "mumm" produces confidence intervals with slightly too small coverage probabilities. On the other hand, method "MAM" produces confidence intervals with very reasonable coverage probabilities.

Figure 6 shows the estimated confidence intervals for the last data scenario. Here we see that the methods produce very similar intervals, which is expected when the scaling effect is not present. The intervals found by the "mumm" method are, though, a bit narrower overall. This is also reflected in Table 11, which shows that the coverage probabilities are slightly too low for these intervals. The table further shows that the rest of the methods produce intervals with proper coverage probabilities, as expected due to the insignificant scaling effect.

4. New F-test for the MAM

In Section 2 it was found that the MAM gives a higher power to detect product differences than the 2-way mixed model, when a scaling effect is present. Using the multiplicative mixed model, however, we obtain an even higher power to detect product differences. In this section, we therefore propose a new and improved F-test for the MAM, for testing the significance of an overall product effect.

When using the 2-way mixed model, the denominator in the F-test, for testing the significance of an overall product effect, is the mean square for the product-by-assessor interaction ($MS_{Interaction}$). However, the interaction con-

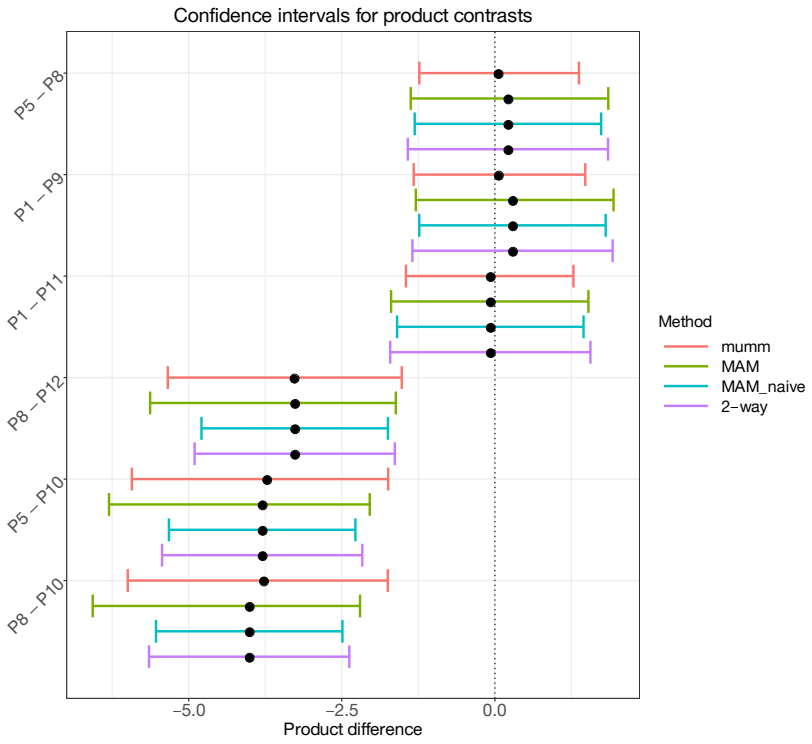


Figure 5: The estimated product contrasts (black dots) and the corresponding 95%-confidence intervals, for the *Colourbalance* attribute. Method "MAM" refers to the procedure suggested in Brockhoff et al. (2015), method "MAM_{naive}" refers to using MAM directly, and method "mumm" refers to the use of R-package *mumm* for finding the profile likelihood based confidence intervals. Only the confidence intervals for the three smallest and the three largest contrasts are plotted.

tains the scale effect, which means that the variation due to the assessors' different use of scale range ends up in the denominator, i.e in the "error". When using the MAM, on the other hand, the denominator is $MS_{Disagreement}$, which is the mean square for the interaction without the mean square for the scale effect. Hence, the scale effect has been removed from the "error" in the F-test, giving an increased power to detect product differences.

In the likelihood ratio test, when using the multiplicative mixed model, the full model is tested against a reduced model, where both the product effect and the scaling effect is removed. This idea is similar to using an F-test for

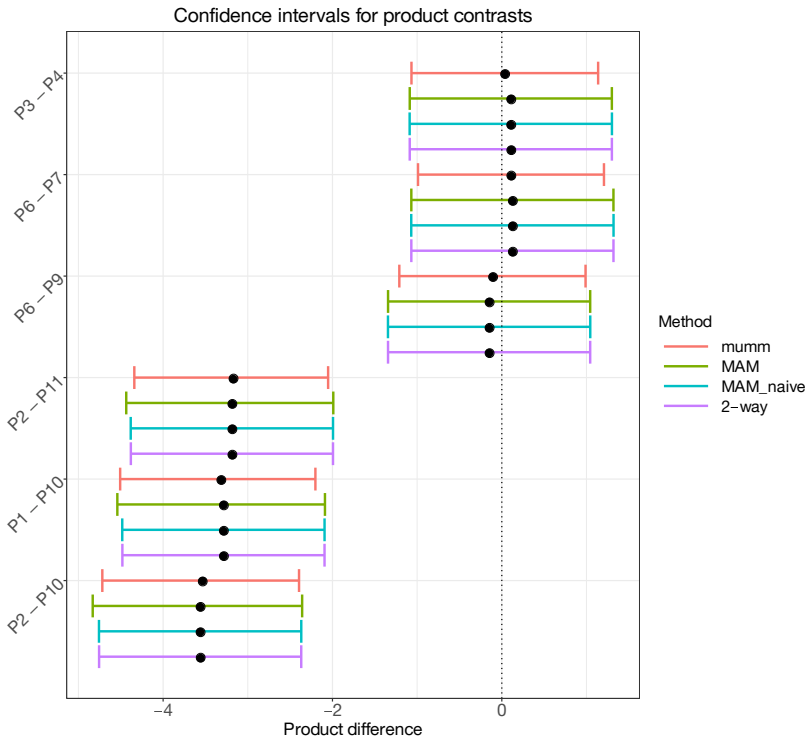


Figure 6: The estimated product contrasts (black dots) and the corresponding 95%-confidence intervals, for the *Depth* attribute. Method "MAM" refers to the procedure suggested in Brockhoff et al. (2015), method "MAM_{naive}" refers to using MAM directly, and method "mumm" refers to the use of R-package *mumm* for finding the profile likelihood based confidence intervals. Only the confidence intervals for the three smallest and the three largest contrasts are plotted.

the MAM with the mean square for the combined effect of product and scaling ($MS_{Product+Scaling}$) in the numerator instead of just $MS_{Product}$. This makes sense, since a large scaling effect is a sign of product difference. Therefore we propose a new F-test for testing the significance of an overall product effect, with the F-statistic:

$$F_{Product} = \frac{MS_{Product+Scaling}}{MS_{Disagreement}} = \frac{(SS_{Product} + SS_{Scaling})/(J + I - 2)}{MS_{Disagreement}}$$

When performing this F-test in the simulation study described in 2.2 and 2.3, we get the results shown in Table 12. Following this, the power to detect product

	Contrast	mumm	MAM	MAM_{naive}	2-way
P5-P8	0.0526	0.9170	0.9370	0.9370	0.9580
P1-P9	0.0567	0.9350	0.9600	0.9600	0.9710
P1-P11	-0.0760	0.9330	0.9650	0.9650	0.9800
P3-P4	0.1229	0.9260	0.9570	0.9570	0.9730
P9-P11	-0.1327	0.9310	0.9540	0.9540	0.9710
P8-P11	-2.9918	0.9170	0.9320	0.8640	0.8960
P5-P12	-3.2264	0.9160	0.9390	0.8470	0.8770
P8-P12	-3.2790	0.9260	0.9380	0.8640	0.8870
P5-P10	-3.7259	0.9180	0.9250	0.8220	0.8580
P8-P10	-3.7785	0.9190	0.9300	0.8260	0.8610

Table 10: The coverage probablities (*Colourbalance*).

differences increases when using the new F-test for the MAM, instead of the original F-test, in the first data scenario. For the second data scenario, the power obtained by the new F-test is even slightly higher than the power of the multiplicative mixed model. For the last data scenario, the use of the new F-test results in a power much lower than the power of the rest of the methods. This is however okay, since this new F-test is intended for data sets, where the scaling effect is present. Therefore we take a closer look at the results of the simulation study for data scenario 1. Figure 7 shows the computed p-values for the 2-way model plotted against the p-values for the new F-test, which looks rather similar to Figure 2, showing the p-values for the 2-way model plotted against the p-values for the multiplicative mixed model. Figure 8 shows the p-values for the new F-test plotted against the p-values for the multiplicative mixed model, which shows that the values seem to follow each other pretty nicely. Thus, using the MAM together with the newly proposed F-test we increase the power to detect product differences, in a similar fashion as when using the multiplicative mixed model together with the likelihood ratio test, when a scaling effect is present.

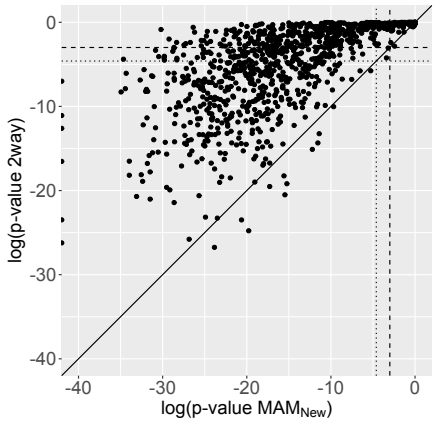


Figure 7: The logarithm of the p-values for the 2-way model plotted against the logarithm of the p-values for the MAM with the new F-test. The dotted and the dashed lines mark p-value = 0.01 and p-value = 0.05, respectively. For the dots on the boundary of the axis the computed value of log(p-value) is $-\infty$.

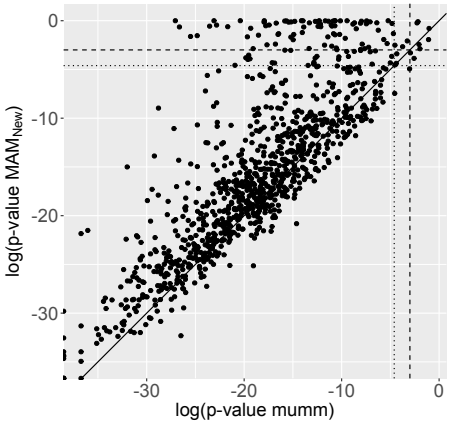


Figure 8: The logarithm of the p-values for the MAM with the new F-test plotted against the logarithm of the p-values for the multiplicative mixed model. The dotted and the dashed lines mark p-value = 0.01 and p-value = 0.05, respectively. For the dots on the boundary of the axis the computed value of log(p-value) is $-\infty$.

	Contrast	mumm	MAM	MAM_{naive}	2-way
P3-P4	0.0331	0.9190	0.9550	0.9550	0.9560
P6-P7	0.1066	0.9330	0.9470	0.9470	0.9500
P6-P9	-0.1089	0.9270	0.9510	0.9500	0.9560
P4-P5	0.1746	0.9210	0.9450	0.9460	0.9470
P3-P5	0.2077	0.9260	0.9550	0.9550	0.9540
P1-P11	-2.9543	0.9540	0.9570	0.9560	0.9580
P8-P10	-3.0139	0.9360	0.9480	0.9360	0.9400
P2-P11	-3.1741	0.9440	0.9490	0.9360	0.9420
P1-P10	-3.3183	0.9320	0.9340	0.9290	0.9340
P2-P10	-3.5382	0.9320	0.9420	0.9290	0.9320

Table 11: The coverage probablities (*Depth*).

		Power	
Data	Product effect	$\alpha = 0.05$	$\alpha = 0.01$
1	$4.80 \cdot 10^{-9}$	91.1	88.0
2	0.0048	79.1	61.4
3	$5.92 \cdot 10^{-2}$	74.5	45.9

Table 12: The results from the simulation studies when using the new F-test for the MAM.

5. Summary and discusion

In this paper we have compared the hypothesis tests for overall product differences when using the MAM, the 2-way mixed model and the multiplicative mixed model. It was found that the use of the multiplicative mixed model results in the highest power to detect product differences, when a scaling effect is present. It was further found, in accordance with Brockhoff et al. (2015), that also the MAM gives an increased power to detect product differences compared to the 2-way mixed model. In addition we found that our novel product differ-

ence F-test for the MAM, results in a power which resembles the power of the multiplicative mixed model.

Through simulation studies, this paper also investigated whether the suggested procedure in Brockhoff et al. (2015), based on the MAM, produce appropriate confidence intervals for product differences. We compared those intervals with the profile likelihood based confidence intervals, based on the multiplicative mixed model, one can obtain from using the R-package *mumm*. It was found that both methods result in proper confidence intervals with reasonable coverage probabilities. Overall, it seems that the MAM based method results in the most reasonable confidence intervals, since the coverage probabilities in some cases were shown to be slightly too low for the profile likelihood based confidence intervals. This might be due to the underestimation of maximum likelihood estimated variance components. It would therefore make sense to maximize the restricted likelihood (REML) instead of the standard likelihood function, but this is, however, not straightforward to do for the multiplicative mixed model. This emphasizes that the MAM based method is a good alternative to the more formal profile likelihood based method.

The results in this work are based on three different data scenarios, with a significant scaling and/or disagreement effect, each with one parameter constellation. In future work, it would be beneficial to investigate how the results generalize to other data sets with other parameter constellations.

5.1. Computation time

The estimation of the parameters in the multiplicative mixed model (3), when fitting the model to the *Dim glass effect* attribute by *mumm*, takes on average 0.36 seconds (average of 100 runs). To achieve the corresponding confidence intervals for all of the 66 product contrasts, 410.54 seconds, i.e. almost 7 minutes, are needed on average[1]. In Pødenphant et al. (2018) the computation

[1]The computer used has 16 GB RAM, an Intel Core i7-6500U processor and runs under the operating system Windows 7 Enterprise. The version of R is 3.4.2.

time for the model estimation when using *mumm* is compared to the computation time for fitting the model by the *NLMIXED* procedure in SAS (Littell et al., 2007), and the authors state that *mumm* is faster than *NLMIXED*. When fitted to the *TVbo* data set, *mumm* is more than 20% faster than *NLMIXED*, and when fitted to larger data sets, the difference can be substantial; In one of their examples, *mumm* is more than 40 times faster. They further note that proper confidence intervals for the multiplicative mixed model cannot be estimated by *NLMIXED*.

The computational burden is, however, immensely reduced for the "MAM" method, with an average computation time of only 0.29 seconds for fitting the MAM and estimating the 66 confidence intervals. This gives a strong advantage to MAM, especially regarding simulation studies and analysis of data from experiments with many products.

5.2. Concluding remarks

The formal modeling approach of fitting the multiplicative mixed model with likelihood methods is advantageous in terms of insight and understanding. Further, it makes it possible to get profile likelihood based confidence intervals. The formal modeling, however, comes with the cost of a relative high computation time. The Mixed Assessor Model is on the other hand very fast to estimate. When a scaling effect is present in the data, both models give an improved power to detect product differences compared to a 2-way mixed analysis of variance, with the multiplicative model resulting in the highest. We have, however, proposed a new F-test for the MAM, which results in a power that resembles, and sometimes even exceeds, the power of the formal modeling approach. Further, we found that the suggested procedure in Brockhoff et al. (2015), based on the MAM, produces appropriate confidence intervals for product differences. In the light of the reduced computation time, we therefore see the MAM as a good alternative to the formal multiplicative mixed model.

6. Compliance with Ethical Standards

Funding: The conduct of the research is fully funded by the Technical University of Denmark (DTU).

Declaration of interest: None

Appendix A. R-code

Listing 1: R-code for estimating product contrast confidence intervals.

```
library(SensMixed)
library(mumm)

#Loading and preparing the data
data = TVbo
data$Product = factor(data$TVset:data$Picture)
data$y = data$Dimglasseffect

#Fitting the MAM
fit_MAM = sensmixed(c("y","y"), c("Product"),c("Assessor"),
                    data = data, product_structure = 1,
                    error_structure = "ONLY-ASS", MAM = TRUE,
                    control = sensmixedControl(calc_post_hoc = TRUE,
                                               MAM_balanced = TRUE,
                                               MAM_adjusted = FALSE))

#Estimating the product contrasts and their confidence intervals
MAM_contrasts = fit_MAM[[5]][,,1][,1]
fit_MAM_posth = fit_MAM[[8]][,,1]
MAM_conf_lower = fit_MAM_posth[,1]
MAM_conf_upper = fit_MAM_posth[,2]

#Fitting the multiplicative mixed model
fit_mumm = mumm(y ~  -1 +  Product + (1|Assessor) +
                  (1|Product:Assessor) +
                  mp(Assessor,Product), data = data)

#Estimating the product contrasts and their confidence intervals
```

```
c1 = combn(12,2)[1,]
c2 = combn(12,2)[2,]
matrix_contrasts = matrix(0,length(c1),nlevels(data$Product))
matrix_contrasts[cbind(1:length(c1),c1)] = 1
matrix_contrasts[cbind(1:length(c2),c2)] = -1
matrix_contrasts_full = cbind(matrix_contrasts,matrix(0,66,5))

mumm_contrasts = matrix_contrasts%*%fit_mumm$par_fix
mumm_conf = confint(fit_mumm, parm = matrix_contrasts_full, level =
    0.95)
```

References

Brockhoff, P. B. (2003). Statistical testing of individual differences in sensory profiling. *Food Quality and Preference*, *14*, 425–434.

Brockhoff, P. B., & Belmonte, F. (2018). Applied univariate statistics. In N. Zacharov (Ed.), *Sensory Evaluation of Sound* chapter 6. CRC Pr I Llc.

Brockhoff, P. B., & Linander, C. B. (2017). Analysis of the data using the r package sensr. In *Discrimination Testing in Sensory Science* (pp. 303–344). Elsevier.

Brockhoff, P. B., Schlich, P., & Skovgaard, I. (2015). Taking individual scaling differences into account by analyzing profile data with the mixed assessor model. *Food Quality and Preference*, *39*, 156–166.

Brockhoff, P. M., & Skovgaard, I. M. (1994). Modelling individual differences between assessors in sensory evaluations. *Food Quality and Preference*, *5*, 215–224.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, *64*, 247–254.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, *70*, 1–21. doi:10.18637/jss.v070.i05.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016a). *lmerTest: Tests in Linear Mixed Effects Models*. URL: `https://CRAN.R-project.org/package=lmerTest` r package version 2.0-32.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2016b). *SensMixed: Analysis of Sensory and Consumer Data in a Mixed Model Framework*. URL: `http://CRAN.R-project.org/package=SensMixed` r package version 2.0-9.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2007). *SAS for mixed models*. SAS institute.

Peltier, C., Brockhoff, P. B., Visalli, M., & Schlich, P. (2014). The mam-cap table: A new tool for monitoring panel performances. *Food Quality and Preference*, *32*, 24–27.

Pødenphant, S., & Brockhoff, P. B. (2016). *mumm: Multiplicative Mixed Models using Template Model Builder*. URL: `https://CRAN.R-project.org/package=mumm` r package version 0.2.0.

Pødenphant, S., Kristensen, K., & Brockhoff, P. B. (2018). The multiplicative mixed model with the mumm r-package as a general and easy random interaction model tool. *arXiv preprint*, .

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: `https://www.R-project.org/`.

Ritz, C., & Skovgaard, I. M. (2005). Likelihood ratio tests in curved exponential families with nuisance parameters present only under the alternative. *Biometrika*, *92*, 507–517.

Smith, A., Cullis, B., Brockhoff, P., & Thompson, R. (2003). Multiplicative mixed models for the analysis of sensory evaluation data. *Food Quality and Preference*, *14*, 387–395.

Vonesh, E. F. (1996). A note on the use of laplace's approximation for nonlinear mixed-effects models. *Biometrika*, *83*, 447–452.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, *80*, 791–795.