Food Quality and Preference 106 (2023) 104792

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Short Communication

Critical evaluation of assessor difference correction approaches in sensory analysis

Justus L. Großmann [a], Johan A. Westerhuis [a,*], Tormod Næs [b], Age K. Smilde [a]

[a] Swammerdam Institute for Life Sciences, University of Amsterdam, Netherlands
[b] Nofima AS, Ås, Norway

ARTICLE INFO

Keywords:
Sensory analysis
Permutation
Product effect significance
Scaling effect
Mixed ANOVA
Mixed assessor model

ABSTRACT

In sensory data analysis, assessor-dependent scaling effects may hinder the analysis of product differences. Romano et al. (2008) compared several approaches to reduce scaling differences between assessors by their ability to maximise the product effect F-values in a mixed ANOVA analysis. Their study on a sensory dataset of 14 cheese samples assessed by twelve assessors on a continuous scale showed that some of these approaches apparently improved the F-value of the product effect. However, this direct comparison is only legitimate if these F-values originate from the same null distribution. To obtain the null distributions of the different correction methods, we employed a permutation approach on the same cheese dataset also used by Romano et al. (2008) and a random noise simulation approach. Based on the empirically obtained null distributions, we calculated the corrected product effect significance to directly compare the performance of the preprocessing methods.

Our results show that the null distributions of some preprocessing methods do not correspond to the expected F-distribution. In particular for the ten Berge method, the null distribution is shifted towards higher F-values. Therefore, an observed increase of the product effect F-value, as compared to the F-value on raw data, does not necessarily lead to increased product effect significance. If p-values are calculated based on such inflated F-values, significance may thus be overestimated. In contrast, calculation of p-values directly from the empirical null distributions obtained by permutation provides a common ground to properly compare method performance. Moreover, we show that differences in reproducibility between assessors, as they exist in real-world sensory datasets, may lead to overestimation of product effect significance by the mixed assessor model (MAM).

1. Introduction

In food science, sensory panels are commonly used to assess the sensory profiles of food products (Amerine, Pangborn, & Roessler, 1965). Despite training and calibration, individual differences between assessors regarding the use of the assessment scale may obfuscate the view on the product profiles. These individual differences manifest themselves in a level effect (differences in assessor means), a scaling effect (differences in product effect magnitude), a disagreement effect (product:assessor interactions not explained by scaling differences) and a variability effect (differences in reproducibility) (Brockhoff, 2003). The handling of these undesired effects has been extensively discussed in literature (Brockhoff & Skovgaard, 1994; Brockhoff, Schlich, & Skovgaard, 2015; Næs, 1990).

Univariate analysis of continuous sensory data is often conducted using a 2-way mixed ANOVA approach with a fixed product effect ν, a random assessor effect a and a random product:assessor interaction g. Sensory assessments for each attribute in a study with I assessors, J products and K replicates can generally be described using an ANOVA model of the following kind:

$$y_{ijk} = \mu + a_i + \nu_j + g_{ij} + \varepsilon_{ijk} \tag{1}$$

where $\varepsilon_{ijk} \sim N(0, \sigma^2)$ and independent (Næs, 1990). Eq. (1) shows that the aforementioned level effect is captured in a, whereas the scaling and disagreement effects will be absorbed by the interaction term g. The F-test for the product effect ν in the mixed ANOVA (Eq. (1)) corresponds to

$$F_\nu = \frac{MS_\nu}{MS_g} \tag{2}$$

with MS_ν and MS_g referring to the mean squares of the product effect and the interaction, respectively (Næs & Langsrud, 1998). Under the null

* Corresponding author.
E-mail address: j.a.westerhuis@uva.nl (J.A. Westerhuis).

https://doi.org/10.1016/j.foodqual.2022.104792
Received 26 August 2022; Received in revised form 8 December 2022; Accepted 20 December 2022
Available online 31 December 2022

hypothesis

$$H_0 : \nu_1 = \cdots = \nu_J = 0 \tag{3}$$

i.e. no difference exists between products, F_ν is expected to be F-distributed with $J-1$ numerator and $(I-1)(J-1)$ denominator degrees of freedom (DFs), if the assumptions about independence, equal variance and normal-distributed residuals hold.

The significance of the product effect is commonly expressed as a p-value (p_ν). The p-value p_ν^* associated with an observed product effect F-value F_ν^* represents the probability of obtaining an F_ν value that is at least as high as the observed F_ν^* value in a situation where H_0 is true. H_0 is rejected for F_ν^* if p_ν^* is lower than the predefined significance level α.

A statistical test is called exact if the probability of a false positive result on H_0 data (type 1 error) is exactly the same as the preset significance level (e.g. $\alpha = 0.05$). A statistical test is termed conservative if this probability of a false positive result on H_0 data is never higher but generally lower than the preset significance level (Good, 2005). In contrast, an anti-conservative test has a probability of producing false positive results that is generally higher than the preset significance level.

Equation (2) shows that absorption of the scaling variance into g_{ij} leads to lower values of F_ν, and thus to a decrease in product effect significance. Therefore, it is desirable to handle the scaling differences between assessors in order to obtain a more significant product effect.

A number of strategies have been devised to account for scaling differences prior to the application of 2-way mixed ANOVA. Romano, Brockhoff, Hersleth, Tomic, and Næs (2008) compared four univariate preprocessing methods to obtain assessor specific scaling factors – standardisation, the *ten Berge* approach, which corresponds to the scaling part of Procrustes rotation (Næs, 1990; Ten Berge, 1977), and two variants of the *assessor model* (Brockhoff & Skovgaard, 1994) – by applying 2-way ANOVA to the scaling-corrected data and assessing the resulting product effect F-values. Their results showed, i.a., that the *ten Berge* method consistently increased F_ν values, delivering the highest F_ν values for 8 out of 12 analysed attributes and seemingly making the product effect for the *fattiness* attribute significant at the $\alpha = 0.01$ significance level.

More recently, Brockhoff et al. (2015) conceived the Mixed Assessor Model (MAM) as an extension to the decomposition in Eq. (1). In the MAM, the interaction term g_{ij} is separated into a disagreement term d_{ij} and an assessor-specific scaling term $\beta_i x_j$:

$$y_{ijk} = \mu + a_i + \nu_j + \beta_i x_j + d_{ij} + \varepsilon_{ijk} \tag{4}$$

The scaling term $\beta_i x_j$ is composed of the centered product averages x_j as a surrogate of the true product effects ν_j and the assessor-specific scaling factors β_i, with $\sum_{i=1}^{I} \beta_i = 0$.

F_ν is calculated as the ratio of MS_ν and MS_d (mean squares of the disagreement effect d) with $J-1$ and $(I-1)(J-1)$ DFs, respectively. Brockhoff et al. (2015) also present two modifications to the MAM: The *conditional MAM* approach (MAM_C) applies Eq. (4) only if the scaling effect is significant at a significance level of, e.g., $\alpha_{scaling} = 0.2$; and the conventional ANOVA (Eq. (1)) is applied otherwise. The *adjusted MAM* (MAM_A) only removes the scaling effects of assessors with positive scaling coefficients from g, as a negative scaling coefficient for a given assessor indicates cardinal disagreement with the majority of assessors. The *conditional adjusted MAM* (MAM_{CA}) combines both approaches, so scaling variance is only removed for assessors with a positive scaling coefficient and if $p_{scaling} < \alpha_{scaling}$. In the modifications of the MAM, the number of denominator DFs corresponds to $(I-1)(J-1) - \iota$, where ι is the number of assessors whose scaling effects were removed.

In this paper, we re-analyse the cheese dataset presented in Romano et al. (2008), focusing on significance of the estimated product effects after applying different assessor difference correction methods. Using a permutation approach, we investigate the null distributions of the different aforementioned methods and check if the actual null distributions correspond to the respective assumed F-distribution. It is essential to calculate product effect significance from the correct null distribution; otherwise, the obtained type I error rates are inaccurate, i.e., the product effect for a given sensory attribute may falsely be deemed significant at a given significance threshold. Thus, obtaining accurate p-values is crucial in the process of deciding if a sensory attribute should be further considered for analysis.

Additionally, we will show for each method how many of the sensory attributes in the given dataset have been improved with respect to product effect significance by correcting for the assessor scaling differences, as compared to applying no correction at all (*raw*). Although this is not a thorough power analysis, it already shows some clear differences between the methods.

2. Permutation strategy

Romano et al. (2008) compared the aforementioned preprocessing methods only by their effect on the size of F_ν. In this study, we set out to investigate the effect of several preprocessing and decomposition methods on ν significance using the same dataset as Romano et al. (2008), by comparing the F_ν values obtained after preprocessing to their respective null distributions. This way, we can benchmark the methods with accurate type I error rates.

In order to obtain the H_0 distributions for the different methods empirically, the case of $H_0 : \nu = 0$ can be simulated repeatedly, recording the observed F_ν values. A straightforward way to achieve this is to randomly permute the values of an attribute in the given sensory dataset. Then, the preprocessing method under investigation is applied to the permuted data, after which the product F-value is calculated according to equation (2). This is repeated for each permutation iteration and the resulting F-values are collected to yield the H_0 distribution. Permutation enables us to mimic the absence of certain effects in the data while ensuring that the resulting data structure resembles the actual dataset. The null distributions obtained through permutation can expose statistical tests that are too optimistic (anti-conservative), and corrected critical values and p-values can be calculated based thereon. As sampling all possible permutations is not computationally feasible, we apply a Monte Carlo permutation approach, where a random subset of all possible rearrangements is considered for analysis (Phipson & Smyth, 2010). Under the assumption that observations are exchangeable, random permutation tests produce asymptotically exact p-values (Good, 2005).

By applying unrestricted permutation, i.e. shuffling the readings of an attribute in a completely random fashion, the product effect ν as well as the assessor effect a and the interaction g vanish, resulting in a data structure that is equivalent to

$$y_{ijk} = \mu + \varepsilon_{ijk} \tag{5}$$

with $\varepsilon_{ijk} \sim N(0, \sigma^2)$ and independent, i.e. the residual error variance σ^2 is equal between assessors. The F_ν null distributions obtained by applying mixed ANOVA (Eq. (1)) on the permuted, preprocessed data will be equivalent to those obtained by performing the same procedure on data sampled directly from a normal distribution $N(0, \sigma^2)$. The F_ν null distributions are expected to be functions of I and J and to be identical between all attributes. A comparable approach has previously been applied to test significance in generalised Procrustes analysis (GPA) (Wu, Guo, De Jong, & Massart, 2002).

However, as also discussed in Romano et al. (2008), the residual error variance of a real dataset may differ between assessors. In that case, it can be argued that the error structure should be preserved by restricting the permutation so that shuffling occurs within assessors only. This leads to a model that includes the mean assessor effect and contains an assessor-specific error variance:

$$y_{ijk} = \mu + a_i + \varepsilon_{ijk} \tag{6}$$

with $e_{ijk} \sim N(0, \sigma_i^2)$, where the F_ν null distribution depends on the σ_i^2 of each assessor and may therefore differ between attributes. This restricted permutation approach has previously been described in the context of sensory data analysis by Xiong, Blot, Meullenet, and Dessirier (2008).

Comparing the two permutation approaches enables us to investigate the effect of neglecting the ANOVA assumption of errors being identically distributed between assessors. While the unrestricted permutation approach conforms with the assumptions of the commonly used mixed ANOVA, the assumption of unequal error variances corresponding to the restricted approach may be more reasonable in reality.

3. Materials and methods

The aforementioned permutation approaches were applied to the sensory dataset used by Romano et al. (2008) which consists of $J = 14$ cheese samples being assessed with regard to 13 sensory properties by $I = 12$ experienced assessors in $K = 2$ replicates, for a total of 336 assessments per attribute. The assessed attributes comprise four odour attributes, eight flavour attributes and *fattiness*. Each attribute was rated on a continuous scale ranging from 1 to 9 with a resolution of 0.1.

Permutation (unrestricted: $2 \cdot 10^5$ randomisations per attribute, restricted: 10^6 randomisations per attribute), preprocessing and subsequent analyses were performed in R 4.0 (R Core Team, 2020), using aov () for fitting the mixed ANOVA and the MAManalysis() function of the SensMixed package (Kuznetsova, Bruun Brockhoff, & Christensen, 2018) for fitting the balanced mixed assessor model (MAM).

To ensure that the findings of the permutation approaches on the given cheese dataset are not just a result of the specific properties of that dataset, we also performed a simulation experiment using randomly generated data. Random data to complement the unrestricted permutation approach was generated according to Eq. (5) by repeatedly sampling from a normal distribution with $\mu = 5$ and $\sigma = 1.5$ in order to produce data that resembles the given sensory dataset (any values for μ and σ are suitable, as long as $\sigma > 0$. Random data to complement restricted permutation was generated according to Eq. (6) by repeatedly sampling I different normal distributions with $\mu + a_i = 5$ and the respective σ_i sampled from a uniform distribution with the interval (0.1,1).

The four preprocessing methods were applied as described in Romano et al. (2008) in order to obtain a set of assessor-specific scaling factors for each attribute. Below, we will briefly describe each of the preprocessing strategies, which are applied separately for each attribute. We also provide `R` code for the scaling methods (see section 6).

The standardisation approach (*std*) standardises each attribute for each individual by subtracting the average value per assessor $(\bar{y}_i)$ and dividing by the standard deviation (s_i) over all products.

$$y_{ijk}^{std} = \frac{y_{ijk} - \bar{y}_i}{s_i} \tag{7}$$

As a result, for each assessor, each attribute has a mean of 0 and a standard deviation of 1.

The *ten Berge* method (*tenb*), which corresponds to the scaling part of the Procrustes rotation method, is discussed in detail in Ten Berge (1977), Næs (1990) and Romano et al. (2008). This method aims to find a scaling factor f_i for each assessor that minimises the sum of squared differences between the scaled assessor profiles for the considered attribute:

$$trace \sum_{i<p} (f_i y_i - f_p y_p)(f_i y_i - f_p y_p)' \tag{8}$$

Each pair of y_i and y_p – the response values of assessors i and p averaged across replicates and zero-centered – multiplied by their optimal scaling factors f_i and f_p should agree as closely as possible.

The assessor model (Brockhoff & Skovgaard, 1994) takes the different use of the scale and different variances ${\sigma_i}^2$ of the assessors into account. The model can be written as

$$y_{ijk} = a_i + \beta_i \nu_j + \varepsilon_{ijk} \tag{9}$$

with $\varepsilon_{ijk} \sim N(0, \sigma_i^2)$, and independent, i.e. the residual error variance σ_i^2 is expected to differ between assessors. Here ν_j represents the product effect, a_i the assessor means, and β_i the different assessor scaling values. The β_i values are estimated in an iterative procedure where individuals are weighted according to their sensitivity.

Romano et al. (2008) suggest two strategies to remove scaling effects based on the $\widehat{\beta}_i$ estimated by the assessor model – a *multiplicative* (*mult*) approach in analogy to standardisation, and an *additive* (*add*) approach. In the multiplicative approach (*mult*), the response data is corrected as follows:

$$y_{ijk}^{mult} = \frac{y_{ijk} - \widehat{a}_i}{\widehat{\beta}_i} \tag{10}$$

However, as discussed in Romano et al. (2008), this approach could lead to biased results because during the iterative estimation of β_i, the individual assessors are weighted differently. The additive approach (*add*) aims to specifically remove the scaling part of the assessor-product interaction,

$$y_{ijk}^{add} = y_{ijk} - (\widehat{\beta}_i - \overline{\widehat{\beta}})\nu_j \tag{11}$$

where $\overline{\widehat{\beta}}$ is the average of all $\widehat{\beta}_i$ values.

The balanced conditional adjusted MAM (MAM_{CA}) was fit with the default settings given in the SensMixed package, i.e. a scaling effect significance threshold of 0.2 and removing only positive scaling effects from the disagreement effect g. The balanced conditional MAM (MAM_C) removes negative scaling effects as well.

P-values were calculated from the empirical cumulative distribution function (ECDF) as the fraction of permutation F_ν values at least as high as the observed F_ν value from the real dataset. In cases where the number of exceedances was below 10, the upper tail of the F_ν null distributions was approximated by a generalised Pareto distribution (GPD) as described in Knijnenburg, Wessels, Reinders, and Shmulevich (2009), using the eva package (Bader & Yan, 2020) for GPD fitting, and the p-value was calculated from the GPD parameters. GPD fitting was carried out with the fitting window starting at the 0.997 quantile, moving up in steps of $1.2 \cdot 10^{-4}$ until a good fit (Anderson-Darling test, $p_{AD} > 0.5$) was obtained.

4. Results and discussion

4.1. Estimating significance by permutation

Application of the unrestricted permutation approach (Eq. (5)) to the sensory dataset from Romano et al. (2008), followed by the different preprocessing methods and mixed ANOVA (Eq. (1)) reveals the F_ν null distributions shown in Fig. 1. With unrestricted permutation, the errors of the considered sensory attribute are pooled across assessors, so that σ_i is equal between assessors. We expect the distributions of the pooled errors to be quite consistent between attributes, as the error distribution for each attribute should approach a normal distribution with an increasing number of assessors, samples and replicates (Liapounoff, 1900).

Hence, the null distributions for the different attributes are in close agreement among each other (Fig. S2) as well as with the null distributions generated from random, normal-distributed data (Fig. S6) with $I = 12$, $J = 14$ and $K = 2$. The F_ν null distributions shown in Fig. 1 are therefore representative for the null distributions of all 13 attributes.

The distributions of F_ν^{raw}, F_ν^{std} and $F_\nu^{MAM_{CA}}$ are in good agreement with the assumed F-distribution with $J - 1$ and $(I - 1)(J - 1)$ DFs. However,

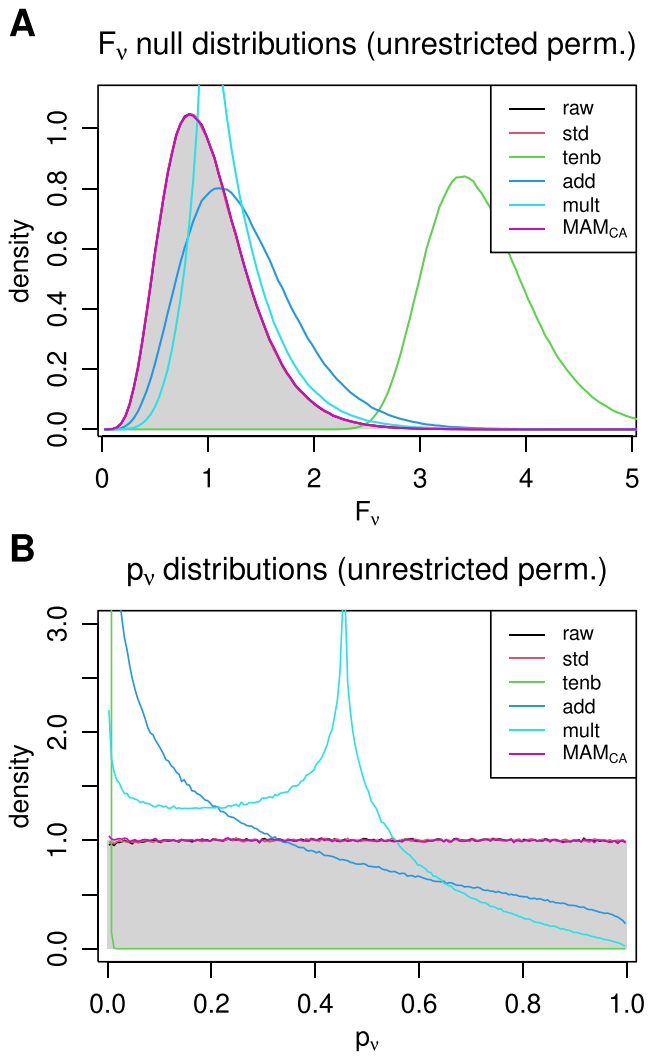

Fig. 1. (A) Null distributions of F_ν after applying different preprocessing and decomposition methods to data generated by repeatedly permuting all 13 attributes of the given cheese dataset, compared to the expected null distribution with J-1 and $(I-1)(J-1)$ DFs (grey). The *raw*, *std* and MAM_{CA} distributions agree with the expected F-distribution. **(B)** Histogram of p-values obtained from the F_ν distributions shown above, assuming that F_ν is F-distributed with $J-1$ and $(I-1)(J\text{-}1)$ DFs. Under *H*0 (no product effect), a uniform distribution of p_ν values is expected (grey). *Raw*: uncorrected data, *std*: standardisation, *tenb*: *ten Berge* method, *add*: additive assessor model, *mult*: multiplicative assessor model, MAM_{CA}: conditional adjusted MAM.

applying mixed ANOVA to data preprocessed with the *ten Berge* (*tenb*) and *assessor model* (*add* and *mult*) methods produces null distributions that are shifted towards higher values of F_ν. Consequently, the p-values calculated from F_ν values produced by these methods assuming that their F_ν null distributions correspond to an F-distribution with $J-1$ and $(I-1)(J-1)$ DFs are underestimated (Fig. 1B).

In other words, this test would be anti-conservative, i.e., product effect significance would be overestimated for *tenb*, *add* and *mult*. To assess the influence of different preprocessing methods on the significance of the product effect, a direct comparison of the respective F_ν values is therefore impossible. Instead, the respective null distributions (as shown in Fig. 1A) should be consulted to calculate p-values that can be used for a proper comparison between methods.

The *ten Berge* preprocessing produces an almost symmetric F_ν null distribution with a mean of 3.59, showing that application of this method will yield a seemingly significant product effect even when applied to permuted or randomly generated data where no product effect is present. For instance, an F_ν value of 3 would be considered highly significant if it arises from uncorrected or standardised data but clearly nonsignificant if it arises from *ten Berge* preprocessed data. This is in line with previous findings that the closely related generalised Procrustes analysis (GPA) will find a consensus in permuted sensory panel data, with the probability of obtaining a consensus by chance increasing with the number of attributes considered (Wakeling, Raats, & MacFie, 1992). Our results show that this is also true for the single-attribute case.

The null distributions of F_ν^{add} and F_ν^{mult} are shifted towards higher values as well. For the given dataset, F_ν^{add} follows an F-distribution with a nonzero noncentrality parameter, while F_ν^{mult} exhibits a sharp peak at 1 and does not resemble an F-distribution.

Using the empirical F_ν null distributions obtained by permutation, critical F-values were calculated for each preprocessing method (Table 1), which show that the critical F_ν values for a typical significance level of $\alpha = 0.01$ differ drastically between methods. Based on the F_ν values obtained from the unpermuted data (Fig. 2A), corrected p-values were calculated directly from the empirical cumulative distribution functions (ECDF) or by fitting a generalised Pareto distribution (GPD) to the distributions' tails (Fig. 2B). The corrected p-values show that mixed ANOVA on the raw (uncorrected) data and on standardised data yielded a significant product effect for 12 out of 13 attributes (*#sign*), while the additive *assessor model* and the MAM_{CA} increased and the *ten Berge* method and the multiplicative *assessor model* decreased the number of significant attributes (Table 1). Standardisation, the additive *assessor model* and the MAM_{CA} improved p_ν for the majority of attributes (*#impr*), whereas the *ten Berge* method and the multiplicative *assessor model* reduced product effect significance. In line with the findings by Romano et al. (2008), the multiplicative *assessor model* preprocessing does not seem appropriate for the given data. Also, although the *ten Berge* method consistently increases F_ν, the comparison to its null distribution reveals that it improves ν significance only for 2 out of 13 attributes.

4.2. *Effect of unequal errors*

In addition to unrestricted permutation, we also performed a restricted (within assessor) permutation approach to account for differences in σ_i between assessors (Eq. (6), Fig. S1). Contrasting the restricted with the unrestricted permutation approach enables us to estimate the effect of disregarding the ANOVA assumption of identically distributed errors. The comparison of null distributions across attributes (Fig. S3) reveals that particularly the MAM_{CA} is sensitive to the error structure of the considered sensory attribute. For some sensory attributes, the F_ν distributions obtained by restricted permutation are shifted to the right, as compared to the expected F-distribution with $J-1$ and $(I-1)(J-2)$ DFs and the distribution obtained by unrestricted permutation. This means that with no product effect present, the F_ν values calculated by the MAM_{CA} on data with unequal σ_i are higher than on data with identical σ_i. In other words, the MAM_{CA} (as well as the additive *assessor model*) may overestimate F_ν in situations where the error is not

Table 1
Critical F_ν-values ($\alpha = 0.01$) for the given dataset were calculated from the ECDF obtained by unrestricted permutation. #sign denotes the number of significant attributes ($\alpha = 0.01$) after application of each method and #impr represents the number of attributes for which p_ν improved by the respective method (as compared to raw, uncorrected data).

Method	$F_{crit}^{\alpha=0.01}$	#sign	#impr
raw	2.26	12	-
standardisation	2.27	12	10
ten Berge method	5.07	8	2
assessor model – add.	2.86	13	10
assessor model – mult.	2.50	5	1
MAM_{CA}	2.27	13	11
$F(J\text{–}1, (I\text{–}1)\,(J\text{–}1))$	2.26	–	–

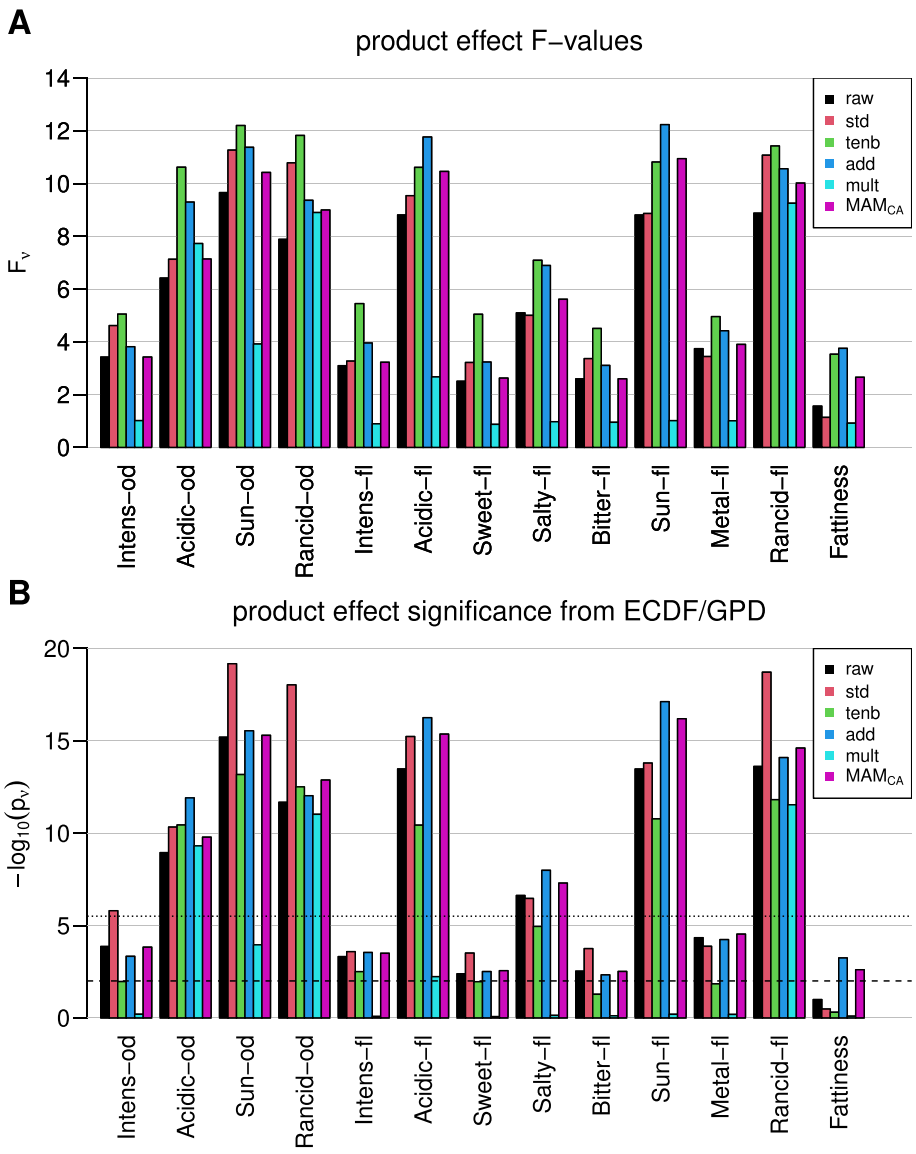

Fig. 2. (A) Product effect F values (F_ν) estimated by mixed ANOVA (Eq. (2)) on the unpermuted cheese data (12 assessors, 14 products, 2 replicates) before (*raw*) and after preprocessing (*std*: standardisation, *tenb*: *ten Berge* method, *add*: additive assessor model, *mult*: multiplicative assessor model, MAM_{CA}: conditional adjusted MAM). Odour attributes are denoted by *–od* and flavours by *–fl*. **(B)** Corrected p-values calculated from the F_ν null distributions generated by unrestricted permutation. P-values greater than $3.125 \cdot 10^{-6}$ (dotted line) were calculated from the ECDF, smaller p-values were calculated from a GPD fit to the distribution tails. The dashed line represents the $\alpha = 0.01$ significance level, which we use to determine whether the product effect for an attribute is significant.

identically distributed between assessors.

Fig. 3 compares the product effect p-values obtained from the unrestricted and restricted permutation null distributions, visualising the effect of taking the presence of unequal σ_i into account. As a consequence of the aforementioned inflation of F_ν for the MAM_{CA} and the additive *assessor model*, the F_ν values observed for the given dataset (see Fig. 2A) are less significant under realistic assumptions (unequal σ_i, vertical axis) than under the idealistic ANOVA assumptions (horizontal axis). In other words, on real data with unequal σ_i, calculating p_ν using the unrestricted permutation null distribution (or assuming an F-distribution with $J-1$ and $(I-1)(J-1)$ DFs, equivalently) will lead to an underestimation of p_ν for the MAM_{CA} and the additive *assessor model* preprocessing, as shown in Fig. 3. In contrast, the standardisation, *ten Berge* and multiplicative *assessor model* null distributions are consistent between permutation methods and attributes, as division by the respective scaling factors SD_i or β_i will evidently equalise the error structures between assessors. For these methods, the size of F_ν will not depend on whether σ_i is unequal between assessors in the considered dataset. However, for the latter two it is essential that their actual F_ν null distributions are used to calculate product effect significance and not the F-distribution with $J-1$ and $(I-1)(J-1)$ DFs.

4.3. Mixed assessor model

A deeper look into the conditional adjusted MAM_{CA} results (Fig. S4) shows that the F_ν null distributions are shifted to the right for attributes that were frequently found to have a significant scaling effect in restricted permutation (Table S1). As the product effect is removed by permutation, we would expect no scaling effect to be present either, $p_{scaling}$ should be uniformly distributed and differences in variance between assessors should be interpreted as differences in reproducibility. Fig. S5 shows that in the case of unrestricted permutation, $p_{scaling}$ is identically distributed between attributes and indeed roughly uniform, whereas in restricted permutation, where σ_i differs between assessors, the MAM_{CA} finds a significant scaling effect more frequently than expected for some attributes. Consequently, the MAM_{CA} removes scaling variance from the interaction effect g more frequently, leading to an inflation of F_ν (see Eq. (2)). The inflation of F_ν is strongest for those attributes whose $p_{scaling}$ distributions are most strongly skewed. The $p_{scaling}$-histograms also show that the underestimation of $p_{scaling}$ and the corresponding inflation of F_ν is not specific to the selected scaling significance cutoff ($\alpha_{scaling}$) of 0.2 for the MAM_{CA} and MAM_C. However, with $\alpha_{scaling}$ approaching zero, the MAM_{CA} and MAM_C become more and more similar to the conventional mixed ANOVA (Eq. (1)), alleviating the

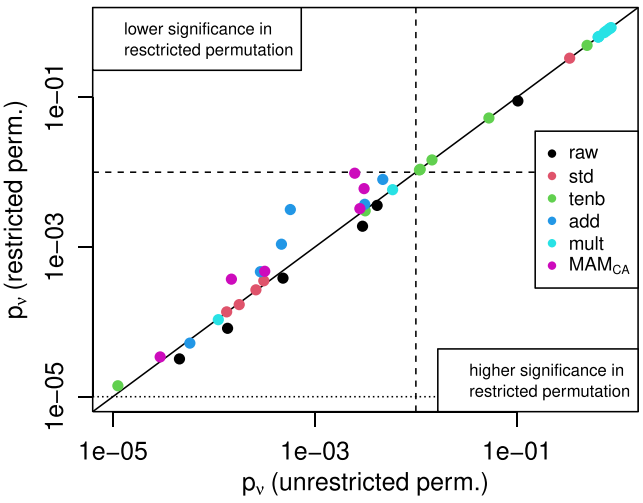

Fig. 3. Comparison of product effect p-values calculated from unrestricted and restricted permutation. The plot only shows the $p > 10^{-5}$ region, where p-values were determined from the ECDF with high confidence. Each dot represents one attribute whose significance was estimated after preprocessing with the method indicated by the fill color. The significance of ν as calculated by the MAM_{CA} and the additive *assessor model* was frequently overestimated by unrestricted permutation, i.e. when differences in σ_i between assessors were disregarded.

issue of an inflated type I error rate possibly at the cost of decreased power in detecting the product effect.

A comparison with the conditional MAM_C reveals that the conditional adjusted MAM_{CA} finds a significant scaling effect (and thus removes the scaling from the interaction effect *g*) less frequently. This means that calculation of the scaling effect in the adjusted manner reduces the inflated chance of obtaining a significant scaling effect with no product effect being present. The F_ν test of the MAM_{CA} is less anti-conservative than that of the MAM_C, which means that the probability of a false-positive result is lower for the MAM_{CA} than for the MAM_C but it can still be higher than the nominal level. Regarding the type I error rate, it is therefore advisable to use the MAM_{CA} rather than the MAM_C on data with unequal errors.

All in all, the results show that the MAM is sensitive to the error structure of the data, because it may interpret unequal σ_i^2 between assessors as scaling effects, even if no product effects are present. This observation can be replicated on randomly generated data (Eq. (6)) without a product effect and with unequal σ_i between assessors (Fig. S6): the F_ν null distribution of the MAM_{CA} is shifted to the right, as compared to the expected F-distribution, implying that the size of the disagreement effect *d* is underestimated. However, if product effect significance is calculated using the respective null distributions generated from restricted permutation, the MAM_{CA} is still the best-performing method on the given dataset.

5. Conclusion

In this work, we revisited and extended the analyses performed by Romano et al. (2008) and presented a permutation-based approach to investigate the statistical validity of the product effect F-test. Our results show that application of certain preprocessing methods to deal with scaling effects in sensory panel data followed by ANOVA will produce F_ν values whose null distributions do not correspond to the expected F-distribution. Consequently, correcting for scaling effects may turn the F-test for the product effect in a mixed ANOVA anti-conservative, which means that the chance of false positive findings is inflated. In particular, the *ten Berge* method produces high F_ν values without a product effect present, which will lead to an overestimation of product effect significance if this fact is not accounted for. An unrestricted and a restricted permutation approach were applied to obtain proper F_ν null distributions and to evaluate the methods under ideal as well as under realistic (σ_i differing between assessors) conditions. Our results show that the error structure of an attribute affects its product effect significance in mixed ANOVA for some preprocessing methods. The permutation results were corroborated by investigations on randomly generated data, showing that our findings generalise beyond the specific dataset used in this study.

Unrestricted permutation produces near-identical null distributions for each sensory attribute, and therefore would need to be performed only once for a given dataset. This approach clearly produces more appropriate type I error rates for the *ten Berge* method and the *assessor model* than just assuming an F distribution with $J - 1$ and $(I - 1)(J - 1)$ DFs. Restricted permutation, which needs to be performed separately for each attribute, additionally accounts for the fact that differences in error variance can result in inflated F_ν values for some methods, such as the mixed assessor model (MAM). Based on the empirical F_ν null distributions obtained through restricted permutation, product effect p-values can be calculated that take each attribute's error structure into account. In conclusion, restricted permutation is more computationally expensive than unrestricted permutation, but delivers statistically more exact results in real-world scenarios where errors are unequal between assessors. Accurate p-values are the foundation for future in-depth power analyses of scaling correction and product effect estimation methods.

6. Data and code availability

The sensory dataset used in this paper as well as the `R` code to perform the scaling correction methods, model fitting and the permutation and simulation approaches have been deposited at github.com/jlgrossmann/scaling-correctionbenchmark.

CRediT authorship contribution statement

Justus L. Großmann: Writing – original draft, Conceptualization, Formal analysis, Visualization. **Johan A. Westerhuis:** Writing – review & editing, Conceptualization. **Tormod Næs:** Methodology, Validation, Supervision. **Age K. Smilde:** Supervision.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [Johan Westerhuis, Justus Grossmann, Age Smilde report financial support provided by the Netherlands Organisation for Scientific Research (NWO Proj. No. 731.015.207).]

Data availability

Data and code have been made available via github: https://github.com/jlgrossmann/scaling-correction-benchmark

Acknowledgement

The authors JLG, JAW and AKS acknowledge funding in the form of a Public-Private Partnership from the Netherlands Organisation for Scientific Research (NWO Proj. No. 731.015.207) in support of this work.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodqual.2022.104792.

References

Amerine, M. A., Pangborn, R. M., & Roessler, E. B. (1965). *Principles of Sensory Evaluation of Food* (1 ed.). Academic Press.

Bader, B., & Yan, J. (2020). *eva: Extreme Value Analysis with Goodness-of-Fit Testing*.

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

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.

Good, P. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses. Springer Series in Statistics* (3 ed.). New York: Springer-Verlag.

Knijnenburg, T. A., Wessels, L. F., Reinders, M. J., & Shmulevich, I. (2009). Fewer permutations, more accurate P-values. *Bioinformatics, 25*, 161–168.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2018). *SensMixed: Analysis of Sensory and Consumer Data in a Mixed Model Framework*. R package version 2.1-0.

Liapounoff, A. (1900). Sur une proposition de la théorie des probabilités. *Bulletin de l'Acadéemie Impéeriale des Sciences de St.-Pétersbourg* 13, 359–386.

Næs, T. (1990). Handling individual differences between assessors in sensory profiling. *Food Quality and Preference, 2*, 187–199.

Næs, T., & Langsrud, Ø. (1998). Fixed or random assessors in sensory profiling? *Food Quality and Preference, 9*, 145–152.

Phipson, B., & Smyth, G. K. (2010). Permutation Pvalues Should Never Be Zero: Calculating Exact Pvalues When Permutations Are Randomly Drawn. Statistical Applications in Genetics and Molecular Biology 9.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Austria: R Foundation for Statistical Computing Vienna.

Romano, R., Brockhoff, P. B., Hersleth, M., Tomic, O., & Næs, T. (2008). Correcting for different use of the scale and the need for further analysis of individual differences in sensory analysis. *Food Quality and Preference, 19*, 197–209.

Ten Berge, J. M. (1977). Orthogonal procrustes rotation for two or more matrices. *Psychometrika, 42*, 267–276.

Wakeling, I. N., Raats, M. M., & MacFie, H. J. (1992). A new significance test for consensus in generalized procrustes analysis. *Journal of Sensory Studies, 7*, 91–96.

Wu, W., Guo, Q., De Jong, S., & Massart, D. L. (2002). Randomisation test for the number of dimensions of the group average space in generalised Procrustes analysis. *Food Quality and Preference, 13*, 191–200.

Xiong, R., Blot, K., Meullenet, J. F., & Dessirier, J. M. (2008). Permutation tests for Generalized Procrustes Analysis. *Food Quality and Preference, 19*, 146–155.