A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data

Benjamin Mahieu, Pascal Schlich, Michel Visalli, Hervé Cardot

Title

A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data

Authors

Benjamin Mahieu[a], Pascal Schlich[a], Michel Visalli[a], Hervé Cardot[b]

[a]Centre des Sciences du Goût et de l'Alimentation, AgroSup Dijon, CNRS, INRAE, Université Bourgogne Franche-Comté, F-21000 Dijon, France.

[b]Institut de Mathématiques de Bourgogne, CNRS, Univ. Bourgogne Franche-Comté, Dijon, France.

Corresponding author:

Benjamin Mahieu; Pascal Schlich

benjamin.mahieu@inrae.fr; pascal.schlich@inrae.fr

Centre des Sciences du Goût et de l'Alimentation, AgroSup Dijon, CNRS, INRAE, Université Bourgogne Franche-Comté, F-21000 Dijon, France.

Highlights

- The usual chi-square framework is not well suited to analyze FC and CATA data
- A modified multiple-response chi-square framework is introduced
- This new framework takes into account the specificities of FC and CATA data
- R package (MultiResponseR) available upon request to the authors and on GitHub

Abstract

Free-Comment (FC) and Check-All-That-Apply (CATA) provide a contingency table containing citation counts of descriptors by products. The analyses performed on this table are most often related to the chi-square statistic. However, such practices are not well suited because they consider experimental units as being the citations (one descriptor for one product by one subject) while the evaluations (vector of citations for one product by one subject) should be considered instead. This results in incorrect expected frequencies under the null hypothesis of independence between products and descriptors and thus in an incorrect chi-square statistic. Thus, analyses related to this incorrect chi-square statistic, which include Correspondence Analysis, can lead to wrong interpretations. This paper presents a modified chi-square square framework dedicated to the analysis of multiple-response data in which experimental units are the evaluations and which is, therefore, better suited to FC and CATA data. This new framework includes a multiple-response dimensionality test of dependence, a multiple-response Correspondence Analysis, and a multiple-response test per cell to investigate which descriptors are significantly associated with which product. The benefits of the multiple-response chi-square framework over the usual chi-square framework are exhibited on real CATA data. An R package called "MultiResponseR" is available upon request to the authors and on GitHub to perform the multiple-response chi-square analyses.

Keywords

- Chi-square statistic
- Multiple-response Correspondence Analysis (MR-CA)
- Multiple-response dimensionality test of dependence
- Multiple-response tests per cell
- Analysis of multiple-response data

1. Introduction

Free-Comment (FC) (ten Kleij & Musters, 2003) and Check-All-That-Apply (CATA) (Adams, Williams, Lancaster, & Foley, 2007) are word citation occurrence-based methods that aim at collecting product descriptions from consumers using either their own words or a mutual predefined list of descriptors. These descriptions are collected without any quantification or product comparison. At the panel level, the collected data constitute count data that are usually stored in a contingency table that contains the number of times each descriptor (in columns) was cited for each product (in rows).

The analysis of these data starts by testing whether overall differences exist between the products. Two approaches can be distinguished to do so. The first one consists of performing a chi-square test while the second one is based on a combination of Cochran's Q statistics (Meyners, Castura, & Carr, 2013). Pursuing the analyses further is only recommended if the existence of overall differences between products is established. In this case, these differences can be visualized using Correspondence Analysis (CA). CA enables to represent the structure of the dependence between products and descriptors on a factorial map that decomposes the whole dependence into axes of maximal and decreasing dependence. As a final step of the analysis, it is important to determine which descriptors are significantly associated with which product. Again, two approaches can be distinguished to do so. The first one is multidimensional alignment (Meyners et al., 2013) that consists of considering a descriptor significantly positively (resp. negatively) associated to a product when their vectors in the sensory space depicted by the CA form an angle lower than or equal to 45° (resp. higher than or equal to 135°). The second approach consists of testing each cell of the contingency table against the null hypothesis of independence using a chi-square test or a Fisher's exact test (Mahieu, Visalli, & Schlich, 2020a; Symoneaux, Galmarini, & Mehinagic, 2012).

All of these approaches but the combination of Cochran's Q statistics are based on the chi-square statistic. The chi-square statistic can be directly used to test for overall differences between the products before performing the CA. The total inertia of CA is the chi-square statistic divided by the grand sum of the contingency table, also called phi-square index. Since multidimensional alignment relies on the CA, it depends also on the chi-square statistic. Finally, the tests per cell approach directly rely on the chi-square statistic since Fisher's exact test can, roughly speaking, be seen as an exact chi-square test.

These common practices assume that all citations are independent experimental units within an evaluation, which is not the case since citations of descriptors by a given subject for a given product are not independent. Instead, one evaluation, i.e. the entire set of descriptors cited by one subject for one product, should be considered as an experimental unit (Loughin & Scherer, 1998). Indeed, considering citations as experimental units implies computing incorrect expected values under the null hypothesis of independence between products and descriptors (Loughin & Scherer, 1998), resulting in an incorrect chi-square statistic. Subsequent analyses of FC and CATA data based on this chi-square statistic are thus also incorrect and can sometimes lead to wrong interpretations.

The present paper aims to overcome the previous limitations by introducing the multiple-response chi-square framework based on the multiple-response chi-square statistic of Loughin and Scherer (1998). This new framework considers experimental units as being the evaluations rather than the citations. First, some notations are introduced and the multiple-response chi-square test of Loughin and Scherer (1998) is presented and adapted to the context of FC and CATA data. Second, the multiple-response Correspondence Analysis (MR-CA) is introduced. Third, the transposition of the methodologies presented in Mahieu et al. (2020a) to the multiple-response chi-square framework is established. Fourth, examples of the benefits of the new framework are given on real CATA data. Finally, an overall discussion and a conclusion are given.

2. Material and methods

2.1. Notations and multiple-response chi-square test of homogeneity

Let us consider an FC or a CATA experiment where S subjects evaluated P products on D descriptors. Each product $p \in \{1, \dots, P\}$ has been evaluated E_p times and the total number of evaluations is equal to $E = \sum_{p=1}^{P} E_p$. Note that in the particular case of balanced experimental design, i.e. when all subjects evaluated all products, then $E = S \times P$. Let us denote by n_{pd} the number of citations of descriptor $d \in \{1, \dots, D\}$ for product p during the E_p evaluations and by C_d the number of citations of descriptor d during all the E evaluations.

Let us denote by π_d^p the probability of descriptor d to be cited for product p. What is under investigation is whether π_d^p differs from one product to another. Using the above notations, the following hypotheses are considered:

$$H_0 : \pi_d^1 = \cdots = \pi_d^P = \pi_d, \qquad \forall d \in \{1, \dots, D\}$$

$$H_A: It\ exists\ d \in \{1, \ldots, D\}\ and\ p, p' \in \{1, \ldots, P\}\ with\ p \neq p' such\ as\ \pi_d^p \neq \pi_d^{p'}$$

Note that this does not correspond to a classical test of homogeneity since, for each product p, multiple descriptors can be selected. Under the null hypothesis, the expected number of citations of descriptors d for product p, denoted by $\boldsymbol{E}(n_{pd})$, is equal to $E_p \times \pi_d$ and can be estimated by $E_p \times {}^{C_d}/_{E}$. The following test statistic, called multiple-response chi-square statistic, is thus introduced:

$$\chi^2_{mr} = \sum_{p=1}^{P} \sum_{d=1}^{D} \frac{\left(n_{pd} - E_p \times {}^{C_d}/_{E}\right)^2}{E_p \times {}^{C_d}/_{E}}$$

As $E_p \times {}^{C_d}/_{E} = E \times \left({}^{E_p}/_{E} \times {}^{C_d}/_{E}\right)$, χ^2_{mr} can also be expressed as:

$$\chi^2_{mr} = \sum_{p=1}^{P} \sum_{d=1}^{D} \frac{\left(n_{pd} - E \times \left({}^{E_p}/_{E} \times {}^{C_d}/_{E}\right)\right)^2}{E \times \left({}^{E_p}/_{E} \times {}^{C_d}/_{E}\right)}$$

As in Loughin and Scherer (1998), it can be shown that the asymptotic distribution of this test statistic under the null hypothesis is complicated because descriptors might not be selected independently. A reasonable option for estimating the distribution of χ^2_{mr} under the null hypothesis is to consider a Monte-Carlo approach (see Section 2.3.1.2).

2.2. The multiple-response Correspondence Analysis

2.2.1. Conceptual difference with the usual Correspondence Analysis for Free-Comment and Check-All-That-Apply data

In usual CA, the products are compared to each other according to their profile. The profile of each product is defined as the proportion of citations of each descriptor for this product relatively to the total number of citations (all descriptors combined) elicited by this same product. Thus, in the context of FC and CATA data, when products elicit different average citation rates (all descriptors combined) then absolute differences in descriptors' citation rates between products are distorted due to this "citation rescaling". The degree of distortion depends on the degree of differences in citation rates between products. For more details on the usual CA, one can refer e.g. to Greenacre (2007). The previous assertions are also applicable to Hellinger-distance-based CA (Rao, 1995; Vidal, Tárrega, Antúnez, Ares, & Jaeger, 2015) because this latter is also based on the products' profiles.

MR-CA overcomes the above limitation by scaling products according to their number of evaluations instead of their number of received citations. It results in comparing products based on their average proportions of citations for each descriptor. This "evaluation scaling" only has importance in the case of unbalanced design. Indeed, products that are more evaluated are likely to elicit more citations of all descriptors and it is necessary to put products on an equal footing before comparing them. To summarize, the propensity of some products to elicit more citations than others does not affect MR-CA while it does with usual CA.

When applied to FC and CATA data, MR-CA can be seen as standing at the frontier between the usual CA of the descriptor by product contingency table and the PCA of the products' average profiles depicted by the descriptors' proportions of citations. MR-CA performs the PCA of the products' average proportions of citations but weighting the descriptors proportionally to their citation rate as in usual CA.

2.2.2. Definition

Similarly, to the usual CA based on the singular value decomposition of the matrix of standardized residuals defined by the usual chi-square statistic, the MR-CA is based on the singular value decomposition of the matrix of standardized residuals defined by the multiple-response chi-square statistic. Using the notations defined in the previous section, let us consider:

- $\boldsymbol{r}$ a column matrix of size $P \times 1$ whose elements equals ${}^{E_p}/_{E}, p \in \{1, \dots, P\}$
- $\boldsymbol{c}$ a column matrix of size $D \times 1$ whose elements equals ${}^{C_d}/_{E}, d \in \{1, \dots, D\}$
- $\boldsymbol{D}_r$ a diagonal matrix of size $P \times P$ whose diagonal elements equal ${}^{E_p}/_{E}, p \in \{1, \dots, P\}$
- $\boldsymbol{D}_c$ a diagonal matrix of size $D \times D$ whose diagonal elements equal ${}^{C_d}/_{E}, d \in \{1, \dots, D\}$
- $\boldsymbol{X}$ a matrix of size $P \times D$ whose general term equal ${}^{n_{pd}}/_{E}, p \in \{1, \dots, P\}, d \in$

$\{1, \dots, D\}$

Using these notations, the MR-CA is based on the singular value decomposition of the matrix $\boldsymbol{S}$ defined as:

$$\boldsymbol{S} = \boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{r}\boldsymbol{c}^t)\boldsymbol{D}_c^{-\frac{1}{2}}$$

Let us denote by $\boldsymbol{U}$ the matrix of left singular vectors of $\boldsymbol{S}$, $\boldsymbol{\Gamma}$ the diagonal matrix of singular values of $\boldsymbol{S}$ and $\boldsymbol{V}$ the matrix of right singular vectors of $\boldsymbol{S}$ such that $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Gamma}\boldsymbol{V}^t$. Similarly to the usual CA, the principal coordinates of the products are defined as $\boldsymbol{D}_r^{-\frac{1}{2}}\boldsymbol{U}\boldsymbol{\Gamma}$ and the so-called contribution coordinates (Greenacre, 2013) of the descriptors are defined as $\boldsymbol{V}$. Note that since this system of coordinates defines a strict biplot as defined in (Gabriel, 1971), it is suggested to use arrows rather than points to display the descriptors' coordinates. This could help practitioners to remember to interpret relations between products and descriptors as scalar products (orthogonal projection) and not “proximities”. Different systems of coordinates could be used for displaying results of MR-CA similarly to usual CA (Greenacre, 2006). However, the one proposed here has two benefits: it enables interpreting maps similarly to Principal Component Analysis (PCA) biplots and the coordinates of the columns (descriptors) reflect their respective contribution to the inertia and to the distances between rows (products) (Greenacre, 2006).

Equivalently, the MR-CA can be defined as the PCA of the matrix $\boldsymbol{D}_r^{-1}\boldsymbol{X}\boldsymbol{D}_c^{-\frac{1}{2}}$. This latter definition of MR-CA better highlights that the distance between two products $p \neq p' \in \{1, \dots, P\}$ in the sensory space depicted by MR-CA called multiple-response chi-square distance is equal to:

$$d_{\chi^2_{mr}}(p, p') = \sqrt{\sum_{d=1}^{D} \frac{E}{C_d}\left(\frac{n_{pd}}{E_p} - \frac{n_{p'd}}{E_{p'}}\right)^2}$$

From the definition of the multiple-response chi-square distance, one can see that the weight given to each product is proportional to its number of evaluations rather than its number of received citations as it is in usual CA. Finally, it should be noted that the number of axes obtained by MR-CA is equal to the minimum between $P - 1$ and D, as in a PCA in which descriptors act as variables and products as individuals, while in usual CA it is equal to the minimum between $P - 1$ and $D - 1$. This difference in the number of axes is because usual CA centers both rows (products) and columns (descriptors) while MR-CA centers only rows.

2.3. Statistical inference for multiple-response chi-square framework

This section transposes the methodologies from Mahieu et al. (2020a) to the multiple chi-square framework.

2.3.1. The dimensionality test of the dependence

2.3.1.1. Conceptual aims for Free-Comment and Check-All-That-Apply data

The aim of this test is twofold. First, it investigates if at least one axis of the MR-CA is significant, that is if some overall differences exist between the products. If no axis is significant, interpreting subsequent analyses including the outputs from MR-CA might lead to over-interpretations. If at least one axis is significant, the second aim of the test is to determine the number of axes that can be considered significant and thus interpreted. Because drawing sensory conclusions based on more than three or four axes can be difficult visually, the number of significant axes is taken into account in subsequent proposed analyses, which are simpler to interpret from a sensory point of view.

2.3.1.2. Technical aspects

It is possible to test if the dependence of each MR-CA axis is significant with a stepwise procedure similarly as for the usual CA (Mahieu et al., 2020a). The idea is to test, at each step k $(k > 1)$, whether the hypothesis of independence between products and descriptors is still rejected while the dependence captured by the axes 1 to $k-1$ was removed. In other words, it is tested if the strength of the dependence is still large enough to be considered significant.

As seen in the previous section, the total number of MR-CA axes, denoted K, is equal to the minimum between $P-1$ and D. Let us consider $\boldsymbol{U}_k$ the matrix of the $K-k+1$ last left singular vectors of $\boldsymbol{S}$, $\boldsymbol{\Gamma}_k$ the diagonal matrix of the $K-k+1$ last singular values of $\boldsymbol{S}$ and $\boldsymbol{V}_k$ the matrix of the $K-k+1$ last right singular vectors of $\boldsymbol{S}$ such that $\boldsymbol{S}_k = \boldsymbol{U}_k\boldsymbol{\Gamma}_k{\boldsymbol{V}_k}^t$. Let us denote by $\chi^2_{mr_k}$ the multiple-response chi-square statistic of the derived contingency table corresponding to the $K-k+1$ last axes of the MR-CA denoted $\boldsymbol{Y}_k$ and defined following the *reconstitution formula* as:

$$\boldsymbol{Y}_k = \left(\boldsymbol{D}_r^{\frac{1}{2}}\boldsymbol{S}_k\boldsymbol{D}_c^{\frac{1}{2}} + \boldsymbol{r}\boldsymbol{c}^t\right) \times E$$

The multiple-response chi-square test associated with the test statistic ${\chi^2_{mr}}_k$ enables testing if the k-th axis of the MR-CA captures a significant dependence between products and descriptors. Note that if $k = 1$ then this test corresponds to the multiple-response chi-square test defined in section 2.1.

The multiple-response chi-square statistic of the products by descriptors contingency table is related to the eigenvalues of the MR-CA by the following equation:

$$\chi^2_{mr} = E \times \sum_{i=1}^{K} \lambda_i$$

where χ^2_{mr} is the multiple-response chi-square statistic of the contingency table, E is the total number of evaluations and λ_i is the i-th eigenvalue of the MR-CA. This relation enables to compute each ${\chi^2_{mr}}_k$ as:

$$\forall k, \qquad {\chi^2_{mr}}_k = E \times \sum_{i=k}^{K} \lambda_i$$

To estimate the distribution of each ${\chi^2_{mr}}_k$ under the null hypothesis, it is proposed to randomly permute the response vectors along products within each subject (Mahieu et al., 2020a; Meyners et al., 2013; Meyners & Pineau, 2010; Wakeling, Raats, & MacFie, 1992; Winkler, Webster, Vidaurre, Nichols, & Smith, 2015), a response vector referring to all citations given for one product by one subject.

To summarize, the dependence between products and descriptors captured by each MR-CA axis can be tested following these steps:

(i) Simulate a large number of contingency tables by randomly permuting the response vectors along products within each subject

(ii) Perform MR-CA on each of the simulated contingency tables

(iii) Compute all ${\chi^2_{mr}}_k^{(*)}$ statistics, $k = 1, \dots, K$, as ${\chi^2_{mr}}_k^{(*)} = E \times \sum_{i=k}^{K} \lambda_i^{(*)}$ for each of the simulated contingency tables

(iv) Compute the p-value of each $\chi^2_{mr_k}$ as the proportion of $\chi^{2\,(*)}_{mr_k}$ under permutation having an equal or a larger value than the observed $\chi^2_{mr_k}$.

2.3.2. Confidence ellipses and discrimination of the products

In MR-CA, as well as in every multivariate analysis providing a product map, superimposing confidence ellipses on product coordinates is crucial to estimate if products are well discriminated. A total bootstrap procedure (Cadoret & Husson, 2013) is proposed to achieve this objective. This procedure consists of generating virtual panels by randomly resampling with replacement the subjects of the actual panel. Then, the product configurations of the virtual panels are rotated on the product configuration of the actual panel thanks to Procrustes rotations. A confidence ellipse is then constructed for each product based on the coordinates of its rotated bootstrap replicates. It is proposed to rely on the significant axes, indicated by the test of dependence presented in section 2.3.1, to determine the number of axes to account for the Procrustes rotations in the total bootstrap procedure.

For each pair of products, to determine if the two products are significantly different, it is proposed to rely on the total bootstrap test (Mahieu, Visalli, Thomas, & Schlich, 2020b) considering the null hypothesis that the two products are not different. For each pair of products, a canonical discriminant analysis based on the rotated bootstrap replicates of the two products is performed. The rotated bootstrap replicates of the two products are then projected on the axis resulting from the canonical discriminant analysis. The distribution of the paired differences of the projected bootstrap replicates is estimated. Finally, the probability of zero to belong to this distribution is estimated and used as a p-value of the test. It is proposed to perform the total bootstrap tests on the significant axes.

2.3.3. Determination of the significant associations between products and descriptors: multiple-response tests per cell

2.3.3.1. Conceptual aims for Free-Comment and Check-All-That-Apply data

These tests aim to investigate the relations between descriptors and products. In particular, they investigate for a given descriptor and a given product if this descriptor is cited for this product in a proportion that significantly differs from the overall average citation proportion of this descriptor all products combined. The tests can be one-sided (positive differences) or two-sided (both positive and negative differences): this choice is up to the discretion of the practitioner. A discussion is given about this choice in Mahieu et al. (2020a).

2.3.3.2. Technical aspects

It is proposed to define a multiple-response test per cell to test the following hypotheses for a given $p \in \{1, \dots, P\}$ and a given $d \in \{1, \dots, D\}$:

$$H_0: \pi_d^p = \pi_d$$

$$H_A: \pi_d^p \neq \pi_d$$

The multiple-response test per cell is based on a Monte-Carlo procedure. In this procedure, for each product $p \in \{1, \dots, P\}$, E_p evaluations are randomly drawn among the subjects having evaluated p and only one evaluation is randomly drawn among each of these subjects. This enables constructing a virtual contingency table under the null hypothesis accounting for both the subject structure of the data and the non-independence of the citations. Indeed, one evaluation is randomly drawn from each subject having evaluated p and one randomly drawn evaluation (that respect the joint distributions of citations of the descriptors) contributes to several cells in the virtual contingency table.

A large number of virtual contingency tables under the null hypothesis can be generated by repeating this procedure. Then, for each cell, the proportion of $n_{pd}^{(*)}$ under the null hypothesis having an equal or a more extreme value than the observed n_{pd} constitute a p-value of the test. The multiple-response tests per cell can be performed with a two-sided alternative hypothesis or a one-sided greater alternative hypothesis.

Finally, it is proposed to perform the multiple-response tests per cell on the derived contingency table corresponding to the significant axes (Mahieu et al., 2020a), denoted $\boldsymbol{Y}_{sig}$, and defined following the *reconstitution formula* as:

$$\boldsymbol{Y}_{sig} = \left(\boldsymbol{D}_r^{\frac{1}{2}} \boldsymbol{S}_{sig} \boldsymbol{D}_c^{\frac{1}{2}} + \boldsymbol{r}\boldsymbol{c}^t\right) \times E$$

Where $\boldsymbol{S}_{sig} = \boldsymbol{U}_{sig}\boldsymbol{\Gamma}_{sig}\boldsymbol{V}_{sig}{}^{t}$ with $\boldsymbol{U}_{sig}$ the matrix of left singular vectors of $\boldsymbol{S}$ corresponding to the significant axes, $\boldsymbol{\Gamma}_{sig}$ the diagonal matrix of singular values of $\boldsymbol{S}$ corresponding to the significant axes and $\boldsymbol{V}_{sig}$ the matrix of right singular vectors of $\boldsymbol{S}$ corresponding to the significant axes.

To perform the multiple-response tests per cell on $\boldsymbol{Y}_{sig}$ rather than on the observed contingency table results in a gain of power without any inflation of the type I error as suggested by the simulation results presented in the Appendix. The simulation results also suggest that the smaller the number of significant axes and the intensity of the dependence between products and descriptors, the higher the gain of power.

2.4. Examples

These examples from two CATA datasets aim to compare outputs obtained from analyses belonging to the usual chi-square framework to those obtained from analyses belonging to the multiple-response chi-square framework. Although these examples deal with CATA datasets, note that the multiple-response chi-square framework is also appropriate to analyze FC data.

2.4.1. Datasets

The datasets are the same from Mahieu, Visalli, Thomas, and Schlich (2021).

The study took place at the Barry Callebaut© Company, Belgium. Seventy regular consumers of milk chocolates (at least once every two weeks) were recruited among the employees of the Barry Callebaut© Company (not implied in sensory and consumer research). They performed a CATA task on four milk chocolates having different recipes: a standard Belgian milk chocolate, a Swiss milk chocolate, a milk compound chocolate, and a protein base milk chocolate. The four products were presented according to a Williams Latin square design. For each product, the CATA task was carried out according to two sensory modalities: texture in the mouth followed by flavor in the mouth. All the CATA descriptors were selected thanks to the expertise of sensory experts from the Barry Callebaut© Company. The collected data were then stored in two contingency tables, one per sensory modality, by cross tabulating the citation counts of the descriptors (columns) by the products (rows).

Since sensory interpretation is out of the scope of this paper dedicated to the comparison of the two chi-square frameworks, the descriptors were renamed *D1*, *D2*, etc. and the products were renamed *P1*, *P2*, *P3*, and *P4*. Finally, for the texture dataset, an additional product called *P5* was artificially created. This product is exactly *P4* except that for *P5* the number of received citations for every descriptor has been divided by two as compared to *P4*. This was made to illustrate the differences between the multiple-response chi-square framework and the usual chi-square framework.

2.4.2. Analyses

All analyses were performed using R 4.0.2 (R Core Team, 2020). The analyses belonging to the multiple-response chi-square framework were performed using the R package "MultiResponseR" developed for this purpose by the authors.

The two contingency tables were analyzed using the following procedure. An alpha risk (Type I error) of 10% was considered as the significance level.

The dimensionality of the dependence between products and descriptors was determined within each chi-square framework using the dimensionality test (2000 simulations) presented in Mahieu et al. (2020a) for the usual chi-square framework and using the dimensionality test (2000 simulations) presented in section 2.3.1 for the multiple-response chi-square framework.

When at least one axis was significant, the corresponding CA (usual or multiple-response) was performed on the contingency table. Outputs of each CA were displayed using a standard biplot (Greenacre, 2013). For each CA, confidence ellipses for the products' coordinates in the sensory space were computed with a total bootstrap procedure using 2000 bootstrap samples. The Procrustes rotations were performed on the significant axes. For each pair of products, a total bootstrap test was performed on the significant axes for assessing the significance of product difference.

For each pair of product and descriptor (cell), a Fisher's exact test was performed for the usual chi-square framework and a multiple-response test per cell as described in section 2.3.3 (2000 simulations) was performed for the multiple-response chi-square framework. All tests per cell were performed with a one-sided greater alternative hypothesis and conducted on the derived contingency table corresponding to the significant axes.

3. Results

Sensory modality	Chi-square framework	Axis 1	Axis 2	Axis 3	Axis 4
Texture	*Usual*	0.447 (<0.001)	0.162 (<0.001)	0.001 (0.9970)	0 (1)
	Multiple-response	0.907 (<0.001)	0.323 (<0.001)	0.079 (<0.001)	0.002 (0.6146)
Flavor	*Usual*	0.243 (<0.001)	0.012 (0.0154)	0.003 (0.0914)	/
	Multiple-response	0.557 (<0.001)	0.089 (<0.001)	0.013 (0.0054)	/

Table 1: Eigenvalues of Correspondence Analysis and corresponding p-values (in brackets) for testing the number of significant axes in the usual and multiple-response frameworks for the two datasets

Table 1 shows that whatever the sensory modality and the axis considered, the eigenvalues of the CA are higher in the multiple-response framework than in the usual one. This suggests that the usual framework underestimates the dependence between products and descriptors. This line of reasoning is reinforced by the example treated by Loughin and Scherer (1998) as they obtained a lower p-value (which is partly a function of the effect size) for their chi-square test in the multiple-response framework than in the usual one. On the dimensionality of the dependence, Table 1 shows that similar conclusions are provided between products and descriptors by the two chi-square frameworks concerning the flavor dataset: three axes capture significant dependence. However, the dependence on the third axis appears more certain (p=0.0054) in the multiple-response chi-square framework than in the usual one (p=0.0914). Concerning the texture dataset, only two axes capture significant dependence within the usual chi-square framework while three axes capture significant dependence within the multiple-response chi-square framework.

Fig. 1 shows that for the texture dataset, the maps depicted by the two first axes of the usual CA (Fig. 1(a)) and the MR-CA (Fig. 1(b)) are very similar: all the products except *P5* and all the descriptors have the same position on the two maps. The only difference between these maps is the location of *P5* being different from *P4* and closer to the origin in MR-CA (Fig. 1(b)) as compared to usual CA (Fig. 1(a)). The reason for this difference lies in the fact that *P4* and *P5* have the same profile (repartition of citations) in the usual CA. On the contrary, the MR-CA captures that *P5* received fewer citations than *P4* for all the descriptors but still following the same pattern of association with the descriptors. This explains the position of *P5* relative to *P4*: *P5* deviates from independence in the same direction that *P4* (same pattern of association with the descriptors) but *P5* is closer to the origin of the coordinates system than *P4* (received fewer citations). Concerning the third significant axis obtained with the multiple-response chi-square framework on the texture dataset (Fig. 1(c)), it mainly traduces that *P5* received fewer citations than *P4* for all descriptors, which is logical. Note that the usual CA is unable to capture this difference between *P4* and *P5*, which explains the non-significance of the third axis for this CA.

For the flavor dataset, Fig. 2 shows that the spaces provided by the usual CA and the MR-CA exhibit different configurations for both products and descriptors. For every descriptor, there is at least one other product that received more citations than *P3*. Thus, in MR-CA, it is associated with no descriptor, which explains its position: *P3* lies at the opposite of every descriptor loadings (Fig. 2(c) & Fig. 2(d)). On the contrary, in usual CA, *P3* seems to be associated with *D1*, *D5,* and *D6* and slightly with *D2* (Fig. 2(a) & Fig. 2(b)). Indeed, in usual CA, the number of citations received by *P3* for every descriptor is rescaled according to its total number of received citations. Thus, the fact that for every descriptor there is at least one other product that received more citations than *P3* is erased in the usual CA. These features of *P3* are the principal explanation of the differences between the spaces provided by MR-CA and usual CA, both applied on the flavor dataset. Another notable difference is between the maps depicted by the two first axes of the usual CA and of the MR-CA (Fig. 2(a) & Fig. 2(c)). On these maps, *P1* and *P2* appear to be more associated with *D1*, *D4,* and *D6* in MR-CA as compared to the usual CA. This difference is due to the opposite phenomenon that occurred with *P3*: *P1* and *P2* received much more citations than *P3* and *P4* for these descriptors and the usual CA distorts this difference while the MR-CA does not.

Concerning the total bootstrap tests, whatever the considered sensory modality and whatever the considered chi-square framework, the conclusions they provided were the same except when considering the pair *P4*/*P5* and the texture dataset. In the usual chi-square framework, *P4* is for sure not different from *P5* ($p = 1$) while *P4* and

P5 are significantly different in the multiple-response chi-square framework ($p < 0.001$). Of course, this is perfectly in line with Fig. 1(a), Fig. 1(b), and Fig. 1(c).

For texture, Fig. 3 shows that differences in the significant associations concern the pairs: *P2-D2*, *P5-D3,* and *P5-D8*. The pair *P2-D2* is significant in the multiple-response and not in the usual chi-square framework because *P2* received more citations of this descriptor than the other products except *P3*. Concerning the product *P5*, it is noticeable that in the usual framework, it is significantly associated with the same descriptors as *P4 (D3, D5, and D8)*, which was expected since *P5* has the same profile that *P4* in this framework. The pair *P5-D8* being significant in the usual framework with a percentage of citations (25.71%) lower than the one of product *P3* (28.57%) which is not significant nicely illustrates the issue of the "citation rescaling" due to considering the citations as experimental units. Since *P3* and *P5* were evaluated the same number of times, it is counterintuitive to have the one with the lowest proportion of citations significant and not the other. However, in the multiple-response framework, both *P3* and *P5* are not significantly associated with *D8*, which is consistent. Regarding the pair *P5-D3*, the association is not significant in the multiple-response chi-square framework while it is in the usual chi-square framework. This difference is due to the "citation rescaling" that occurs in the usual chi-square framework and not in the multiple-response one.

Concerning the flavor dataset, several differences are shown in Fig. 3 between the conclusions provided by the two chi-square frameworks on descriptor by product significant associations. As was suggested by Fig. 2, *P1* and *P2* are significantly associated with *D1*, *D4,* and *D6* in the multiple-response chi-square framework while only *P2* is significantly associated with only *D4* and *D6* in the usual chi-square framework. This difference is because *P1* and *P2* received much more citations than *P3* and *P4* for these descriptors. On the contrary, without the "citation rescaling", since *P3* and *P4* received fewer citations, they got less significance in the multiple response framework; precisely, *P3-D5* and *P4-D3* are no longer significant in this framework. Finally, it is noticeable that the counterintuitive conclusion in the usual chi-square framework on the significant association of *D3* with *P4* and not with *P1* and *P2* while these received a higher percentage of *D3* citations than *P4*, no longer holds in the multiple-response chi-square framework.

4. Discussion

To the best of our knowledge, it is the first time that a chi-square framework properly taking into account multiple-response data is introduced. The proposed analyses including the test of dimensionality, the product confidence ellipses, the pairwise product comparisons, and the product by descriptor association tests, the three of them being conducted on the significant axes, are all originals. This multiple-response chi-square framework fits perfectly to FC and CATA data. However, this multiple-response chi-square framework is not restricted to be used only in sensory and consumer science and can be used to analyze any multiple-response data whatever the field they come from.

The examples presented in this paper showed that the multiple-response chi-square framework is better suited than the usual chi-square framework to analyze FC and CATA data. A major benefit of using the multiple-response chi-square framework is that when the experimental design is balanced, every product is equally weighted. This is more appropriate and leads to logical outputs as opposed to the usual chi-square framework that can lead to counterintuitive outputs. Indeed, it sounds more logical to weight the products equally and not rescale them according to their number of received citations when they have been evaluated the same number of times. Note that an equivalent weighting of the products using the usual chi-square framework is almost impossible since products are very unlikely to receive the same number of citations at the panel level. The multiple-response tests per cell introduced in this paper take into account all the specific aspects of FC and CATA data, especially the non-independence of citations between descriptors.

The conclusions provided by the two chi-square frameworks are not always necessarily different. For example, they would have been almost the same on the texture dataset if *P5* had not been artificially added to the dataset. The more different the citation rates (all descriptors combined) between products are, the more the conclusions drawn from the usual chi-square framework will differ from the multiple-response one. The products likely receive different numbers of citations when some products have few sensory characteristics while some others have a lot or when some products present obvious characteristics while the characteristics of the other products are more subtle; these kinds of situations are likely to occur in sensory evaluation.

Since the multiple-response chi-square framework relies heavily on Monte-Carlo and bootstrap simulations, the results of the proposed analyses are not instantaneous. For the datasets used as examples, it took around 30 seconds by dataset to obtain the results of all analyses. However, this computation time increases with the number of evaluations and thus with the number of subjects and products. For large datasets (e.g. 3000 evaluations), it takes around 5 minutes to obtain the results using the settings of this paper.

5. Conclusion

For the analysis of Free-Comment and Check-All-That-Apply data, the paper proposes to replace the usual chi-square framework with a new multiple-response chi-square framework taking into account dependence among citations within an evaluation. It is thus statistically valid while the former was not. The new framework includes a test of dimensionality, a Correspondence Analysis with confidence ellipses, a test for pairwise product comparison, and a test of significance of product by descriptor associations. Note that ellipses, tests of product comparisons, and tests of association with descriptors are the three of them computed on the significant axes of dependence. The basic difference introduced by this new framework is not to longer consider citations (one descriptor for one product by one subject) as experimental units, but to rely on evaluations (vector of citations for one product by one subject) as being the experimental units. Simulations showed that testing the significance of product by descriptor associations on the significant axes of dependence increased power in detecting product by descriptor associations without any inflation of the type I error. The new approaches are supported by an R package called “MultiResponseR” and available upon request to the authors and on GitHub.

Appendix: Simulations

To investigate the benefits and/or the downsides of performing the multiple-response tests per cell on the derived contingency table corresponding to the significant axes, simulations of sensory data were performed. For every simulation, 60 subjects, 5 products, and 10 descriptors were considered. The 5 products were considered as being evaluated by the 60 consumers, as it is common in sensory evaluation. The descriptors marginal probabilities were randomly chosen and were the following: 0.20, 0.56, 0.26, 0.23, 0.21, 0.30, 0.20, 0.42, 0.52, 0.75. From these marginal probabilities, the matrix of expected probabilities under the null hypothesis of independence between products and descriptors was computed. This matrix contained 50 cells (5 products × 10 descriptors).

Some deviation from independence was then added iteratively to these expected probabilities such that at each step, one axis of dependence was added orthogonally to the previous axes. On the first added axis, two products were differentiated on six descriptors. On the second added axis, two products were differentiated on four descriptors. On the third added axis, two products were differentiated on two descriptors. On the fourth added axis, four products were differentiated on four descriptors. This enabled to control the true dimensionality of the dependence between products and descriptors. The cells that deviated from the null hypothesis did with equal intensity but with opposite direction to keep the marginal probabilities fixed. Two levels of deviation intensity were considered: 0.1 and 0.2. 8 matrices (4 levels of dimensionality × 2 levels of deviation intensity) of probabilities were thus generated. Each of the 8 matrices contains 50 cells (5 products × 10 descriptors).

For each of these 8 matrices, 1000 datasets were simulated. Each of these datasets was generated by adding 60 individual data (the subjects). Each individual data was generated by performing a random Bernoulli draw for each of the 50 cells according to the specified probability given in the matrix.

For each of the 8000 datasets (8 matrices of probabilities × 1000 generated datasets), the number of significant axes was considered unknown and was determined using the dimensionality test presented in section 2.3.1. The multiple-response tests per cell were then performed on either the observed table or the derived contingency table corresponding to the significant axes returned by the test. The p-values of the multiple-response tests per cell were stored.

For each combination of the factors deviation intensity (0.1 or 0.2), dimensionality (one axis, two axes, etc.), and table (observed or derived) and for each of the 50 cells, the proportion of test (among the 1000 datasets) rejecting the null hypothesis was computed at the following nominal alpha risks: 5%, 7.5%, and 10%. Then, the results from a given cell were assigned either to the group H0 if its probability was not modified or to the group H1 otherwise. Finally, the average proportion of rejection of the null hypothesis was computed within each group (H0 or H1), number of dimensions, and deviation intensity. The results are presented in Table 3.

Deviation intensity	Dimensionality	Nominal alpha risk = 5%				Nominal alpha risk = 7.5%				Nominal alpha risk = 10%			
		H0 derived table	H0 observed table	H1 derived table	H1 observed table	H0 derived table	H0 observed table	H1 derived table	H1 observed table	H0 derived table	H0 observed table	H1 derived table	H1 observed table
	1	0.020	0.034	0.521	0.434	0.030	0.052	0.592	0.507	0.040	0.071	0.644	0.562
0.1	*2*	0.029	0.034	0.461	0.444	0.044	0.051	0.537	0.514	0.061	0.069	0.595	0.569
	3	0.032	0.032	0.451	0.450	0.049	0.049	0.523	0.519	0.069	0.069	0.582	0.577
	4	0.034	0.032	0.532	0.536	0.052	0.049	0.594	0.599	0.070	0.068	0.643	0.646
	1	0.018	0.032	0.987	0.955	0.027	0.050	0.991	0.969	0.037	0.069	0.994	0.978
0.2	*2*	0.028	0.033	0.973	0.960	0.041	0.050	0.982	0.973	0.058	0.068	0.988	0.980
	3	0.030	0.032	0.966	0.962	0.046	0.048	0.977	0.973	0.064	0.066	0.984	0.981
	4	0.029	0.030	0.973	0.974	0.046	0.046	0.982	0.982	0.066	0.066	0.987	0.987

Table 3: Average proportion of rejection of the null hypothesis among the 1000 simulations depending on the deviation intensity, the dimensionality, the nominal alpha risk, the table considered, and the deviation from the null hypothesis or not.

Table 3 shows that the empirical type I error never exceed the nominal alpha risk in group H0 for both approaches, which suggests that both approaches are valid. It can be seen that the empirical type I error in the H0 group was even slightly lower when considering the derived table which is a nice feature.

The percentage of rejections in group H1 (estimating test power) was higher when considering the derived table as compared to the observed table whatever the combination of factors considered except with a dimensionality of 4. Therefore, performing the multiple-response tests per cell on the derived contingency table corresponding to the significant axes enables gaining power without increasing type I error. It should also be noted that the smaller the dimensionality of the dependence, the higher the gain of power. It is logical because a low dimensionality maximizes the difference between the derived table and the observed one. Finally, it should also be noted that the gain in power is higher with the lower independence deviation (0.1 vs 0.2), that is with the more complex/subtle situation. This is a nice feature arguing in favor of performing the multiple-response tests per cell on the derived contingency table corresponding to the significant axes.

Acknowledgments

This study is part of a Ph.D. financed by the Region Bourgogne-Franche-Comté and the SensoStat Company.

The authors would like to thank Barry Callebaut© for providing the chocolate samples.

Calculations in the Appendix were performed using HPC resources from PSIUN CCUB (Centre de Calcul de l'Université de Bourgogne, France). The Institut de Mathématiques de Bourgogne is supported by the EIPHI Graduate School (contract ANR-17-EURE-0002).

References

Adams, J., Williams, A., Lancaster, B., & Foley, M. (2007). Advantages and uses of check-all-that-apply response compared to traditional scaling of attributes for salty snacks. In, *7th Pangborn Sensory Science Symposium*. Minneapolis, USA.

Cadoret, M., & Husson, F. (2013). Construction and evaluation of confidence ellipses applied at sensory data. *Food Quality and Preference, 28*(1), 106-115.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika, 58*(3), 453-467.

Greenacre, M. (2006). Tying up the loose ends in simple, multiple, joint correspondence analysis. In, *Compstat*. Heidelberg: Physica-Verlag HD.

Greenacre, M. (2007). *Correspondence Analysis in Practice, second edition*.

Greenacre, M. (2013). Contribution Biplots. *Journal of Computational and Graphical Statistics, 22*(1), 107-122.

Loughin, T. M., & Scherer, P. N. (1998). Testing for Association in Contingency Tables with Multiple Column Responses. *Biometrics, 54*(2), 630-637.

Mahieu, B., Visalli, M., & Schlich, P. (2020a). Accounting for the dimensionality of the dependence in analyses of contingency tables obtained with Check-All-That-Apply and Free-Comment. *Food Quality and Preference, 83*.

Mahieu, B., Visalli, M., Thomas, A., & Schlich, P. (2020b). Free-comment outperformed check-all-that-apply in the sensory characterisation of wines with consumers at home. *Food Quality and Preference, 84*.

Mahieu, B., Visalli, M., Thomas, A., & Schlich, P. (2021). An investigation of the stability of Free-Comment and Check-All-That-Apply in two consumer studies on red wines and milk chocolates. *Food Quality and Preference, 90*.

Meyners, M., Castura, J. C., & Carr, B. T. (2013). Existing and new approaches for the analysis of CATA data. *Food Quality and Preference, 30*(2), 309-319.

Meyners, M., & Pineau, N. (2010). Statistical inference for temporal dominance of sensations data using randomization tests. *Food Quality and Preference, 21*(7), 805-814.

R Core Team. (2020). R: A language and environment for statistical computing. In. Vienna, Austria: R Foundation for Statistical Computing.

Rao, C. R. (1995). A Review of Canonical Coordinates and an Alternative to Correspondence Analysis Using Hellinger Distance. *Questiio, 19*, 23-63.

Symoneaux, R., Galmarini, M. V., & Mehinagic, E. (2012). Comment analysis of consumer's likes and dislikes as an alternative tool to preference mapping. A case study on apples. *Food Quality and Preference, 24*(1), 59-66.

ten Kleij, F., & Musters, P. A. D. (2003). Text analysis of open-ended survey responses: a complementary method to preference mapping. *Food Quality and Preference, 14*(1), 43-52.

Vidal, L., Tárrega, A., Antúnez, L., Ares, G., & Jaeger, S. R. (2015). Comparison of Correspondence Analysis based on Hellinger and chi-square distances to obtain sensory spaces from check-all-that-apply (CATA) questions. *Food Quality and Preference, 43*, 106-112.

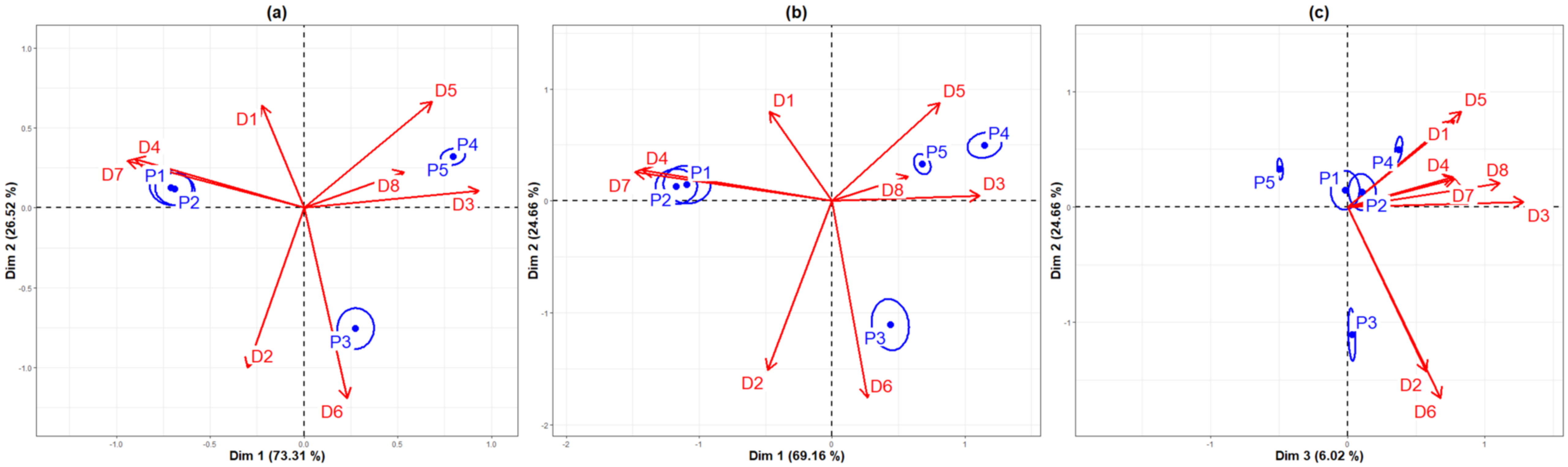
Wakeling, I. N., Raats, M. M., & MacFie, H. J. H. (1992). A New Significance Test for Consensus in Generalized Procrustes Analysis. *Journal of Sensory Studies, 7*(2), 91-96.

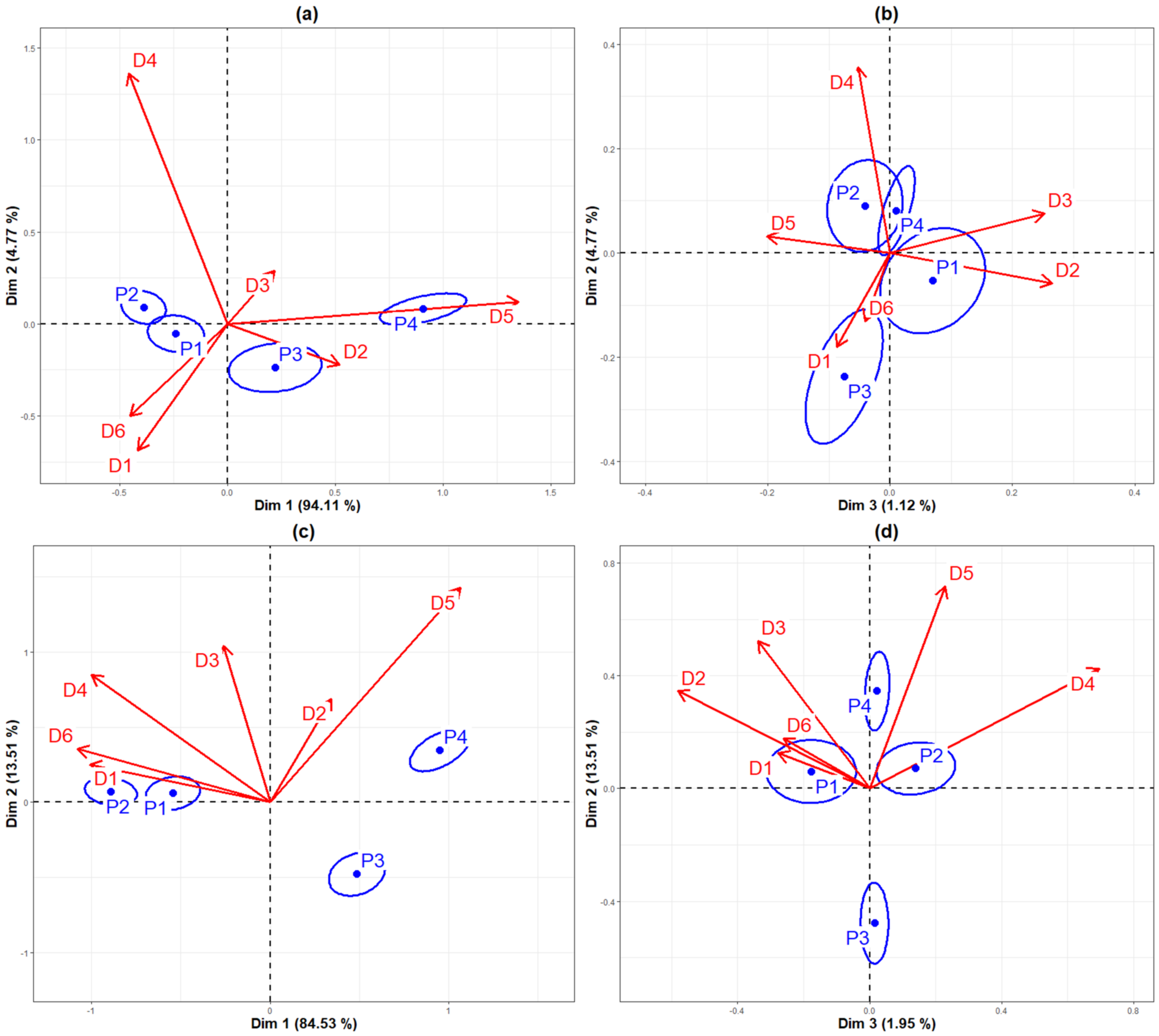
Winkler, A. M., Webster, M. A., Vidaurre, D., Nichols, T. E., & Smith, S. M. (2015). Multi-level block permutation. *NeuroImage, 123*, 253-268.

Fig. 1: Biplot from Correspondence Analysis of the texture dataset: (a) usual CA (axes 1-2), (b) MR-CA (axes 1-2) and (c) MR-CA (axes 3-2).

Fig. 2: Biplot from Correspondence Analysis of the flavor dataset: (a) usual CA (axes 1-2), (b) usual CA (axes 3-2), (c) MR-CA (axes 1-2) and (d) MR-CA (axes 3-2).

Fig. 3: Descriptors by product percentages of citations across the panel. Highlighted cells denote a significant (α = 10%) Fisher exact tests per cell in the usual chi-square framework or a significant (α = 10%) multiple-response test per cell (2000 simulations) in the MR chi-square framework


(a)
(b)
(c)
Dim 1 (73.31 %)
Dim 2 (26.52 %)
Dim 1 (69.16 %)
Dim 2 (24.66 %)
Dim 3 (6.02 %)
Dim 2 (24.66 %)
D1
D2
D3
D4
D5
D6
D7
D8
P1
P2
P3
P4
P5



(a)
D4
D3
P2
P1
P4
D5
P3
D2
D6
D1
Dim 1 (94.11 %)
Dim 2 (4.77 %)
(b)
D4
P2
P4
D3
D5
P1
D2
D6
D1
P3
Dim 3 (1.12 %)
Dim 2 (4.77 %)
(c)
D5
D3
D4
D2
D6
P4
D1
P2
P1
P3
Dim 1 (84.53 %)
Dim 2 (13.51 %)
(d)
D5
D3
D2
P4
D4
D6
P2
D1
P1
P3
Dim 3 (1.95 %)
Dim 2 (13.51 %)

		Usual					Multiple-response				
		P1	P2	P3	P4	P5	P1	P2	P3	P4	P5
Texture	D1	24.29	28.57	0	18.57	9.29	24.29	28.57	0	18.57	9.29
	D2	24.29	27.14	44.29	1.43	0.71	24.29	27.14	44.29	1.43	0.71
	D3	11.43	10	50	81.43	40.71	11.43	10	50	81.43	40.71
	D4	61.43	62.86	10	4.29	2.14	61.43	62.86	10	4.29	2.14
	D5	2.86	1.43	4.29	34.29	17.14	2.86	1.43	4.29	34.29	17.14
	D6	11.43	12.86	57.14	15.71	7.86	11.43	12.86	57.14	15.71	7.86
	D7	61.43	65.71	10	2.86	1.43	61.43	65.71	10	2.86	1.43
	D8	12.86	20	28.57	51.43	25.71	12.86	20	28.57	51.43	25.71
Flavor	D1	68.57	74.29	27.14	18.57		68.57	74.29	27.14	18.57	
	D2	7.14	1.43	4.29	12.86		7.14	1.43	4.29	12.86	
	D3	37.14	34.29	12.86	31.43		37.14	34.29	12.86	31.43	
	D4	32.86	51.43	4.29	10		32.86	51.43	4.29	10	
	D5	7.14	2.86	15.71	50		7.14	2.86	15.71	50	
	D6	74.29	81.43	27.14	20		74.29	81.43	27.14	20	