Sensory analysis with consumers using Free-Comment : analyses, performances and extensions

Benjamin Mahieu

THESE DE DOCTORAT DE L'ETABLISSEMENT UNIVERSITE BOURGOGNE FRANCHE-COMTE

PREPAREE A :

Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement,
Unité Mixte de Recherche : Centre des Sciences du Goût et de l'Alimentation

Ecole doctorale n°554 : Environnements – Santé

Doctorat de Sciences de l'Alimentation

Sensory analysis with consumers using Free-Comment: analyses, performances and extensions

Par M. Benjamin MAHIEU

Thèse présentée et soutenue à Dijon le 15/10/2021

Composition du Jury :

Mme Claire SULMONT-ROSSE	Directeur de recherche, INRAE	Président
M. Hervé ABDI	Professeur, Université du Texas	Rapporteur
Mme Anne SAINT-EVE	Maitre de conférence HDR, AgroParisTech	Rapporteur
M. Philippe COURCOUX	Maitre de conférence, ONIRIS	Examinateur
M. Ronan SYMONEAUX	Chargé de recherche HDR, ESA	Examinateur
M. Pascal SCHLICH	Directeur de recherche, INRAE	Directeur de thèse
M. Hervé CARDOT	Professeur, Université de Bourgogne	Co-directeur de thèse
M. Michel VISALLI	Ingénieur, INRAE	Co-encadrant de thèse
M. Arnaud THOMAS	Cadre scientifique, SensoStat	Invité

"*If we knew what it was we were doing,*
it would not be called research, would it?"

Albert Einstein

Abstract

Free-Comment (FC) consists in panelists describing the products using their own terms. Despite its benefits, notably the circumvention of limitations inherent to pre-established lists of sensory descriptors, FC remains rarely used because its performances are not well documented and its analyses and range of application remain limited. This thesis aims to overpass these limitations, highlighting the benefits and the potency of FC and thus put it in the spotlight for sensory analysis with consumers.

For the pretreatment of FC data, a semi-automatized procedure is proposed. It enables the practitioners to extract an *a posteriori* list of sensory descriptors with a compromise between minimizing the loss of information and maximizing the quickness of the pretreatment. For the statistical analysis of FC data, operating in the significant subspace of product by sensory descriptor dependences is proposed together with the multiple-response chi-square framework that better takes into account the structure of the pretreated data than the usual chi-square framework. These analyses have been implemented into a R-package downloadable from GitHub.

The performances of FC have been compared to those of Check-All-That-Apply (CATA), the most popular method for descriptive sensory analysis with consumers. Two performance criteria have been investigated: the discrimination power and the stability of the product characterization. Regarding both criteria, FC turned out to perform as well as CATA, if not better.

Two extensions of FC are proposed. The first one, Free-Comment Attack-Evolution-Finish (FC-AEF), directs the descriptions towards the temporal aspect of the sensory perception. The second one, Ideal-Free-Comment (IFC) paired with liking scoring, identifies the drivers of liking and characterizes the ideal product thanks to FC. An application of these two methods was carried out, demonstrating their ability to fulfill their aims.

Overall, this work demonstrated the potency and the versatility of the FC method. It opens new perspectives for sensory analysis with consumers and it should promote a larger use of FC in that field.

Keywords: Open-ended questions; Free-Comment; Sensometrics; Sensory analysis; Consumer studies

Acknowledgments

I would like to thank first the Région Bourgogne-Franche-Comté and the SensoStat Company for having initiated the adventure by funding this Ph.D. as well as Pascal Schlich and Hervé Cardot for giving me their trust and the opportunity to live this wonderful adventure the Ph.D. is.

Thanks to the different partners having accepted to provide their products to support my works.

Thanks to the teams of the Center for Taste and Feeding Behavior for their welcome during these three years. Particular thanks to Christine Chabert for its availability and efficiency in helping me through administrative processes.

Thanks to Hervé Abdi and Anne Saint-Eve for having accepted to report this Ph.D. thesis. Thanks to Phillippe Courcoux, Claire Sulmont-Rossé and Ronan Symoneaux for having accepted to form the board of this Ph.D.

Thanks to Phillippe Courcoux and Ronan Symoneaux for having accepted to form my Ph.D. committee as well as for their benevolence and the wise pieces of advice they provided me.

I warmly thank the members of the ChemoSens team for their welcome during these three years. Thanks to Betty Hoffarth, Catherine Pédron, Sylvie Cordelle and Anne-Laure Loiseau for having supported me during the experimentation phases and guided me through their administrative aspects. Thanks to Christine Lange for the interesting discussions we had during coffee breaks and meals.

Huge thanks to Hervé Cardot for having pushed me to go deeper into statistical reasoning as this gave rise to one work of this Ph.D. Thanks to him for having contributed to render my reflection more rigorous.

Huge thanks to Arnaud Thomas for having followed the development of this Ph.D. closely. Thanks to him for the wise pieces of advice he provided me and for his external point of view that helped me several times. Thanks to him as well for the less scientific but equally interesting discussions we had.

Particular huge thanks to Michel Visalli for his day-to-day supervision, guidance, supporting and pieces of advice. Thanks to him for the countless interesting discussions we had especially those having given rise to several works of this Ph.D. Thanks to him for these coffee breaks spent at remaking the world, they largely contributed to making these three years pleasant.

Particular huge thanks to Pascal Schlich, the Ph.D. advisor I wish any Ph.D. student had the luck to have. Thanks to him for having given me his trust and let me conduct my Ph.D. in the direction I wanted to. Thanks to him for having guided me, helped me valorizing my works, and for his constant support. Thanks to him for the long interesting discussions we had and his wise pieces of advice. Thanks to him for the courses about the history of sensory analysis

and statistics. Thanks to him for having shared his knowledge and experience with me.

Thanks to my two cats who were great office colleagues during the two lock-downs.

Thanks to my friends for having asked me "Hey, does your Ph.D. go well?" every time we saw each other. Thanks to them for pretending to be interested while still turning every sentence into something funny when I started developing my works.

Thanks to my family for their constant support. Thanks to my grandmother for telling me regularly "One day, you should explain to me what you are doing at work!". Thanks to my sister who is the only person in my family I am confident she understands every ins and outs of a Ph.D. A particular warm thanks to my parents for their flawless support at every moment of my life.

Finally, I would like to give a huge thanks to my number one support, the woman sharing my life, Victoire, who accepted to follow me into this adventure, and without who my life would not be as wonderful.

"*Don't let the sun go down without saying thank you to someone, and without admitting to yourself that absolutely no one gets this far alone.*"

Stephen King

Foreword

This Ph.D. thesis has been possible thanks to a co-financing from the Region Bourgogne-Franche-Comté and the SensoStat Company and has received the thesis award from the French Society of Sensory Analysis in 2021.

Scientific valorizations

Publications

Mahieu, B., Visalli, M., & Schlich, P. (2020). 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. (2020). 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. (2020). Using Free-Comment with consumers to obtain temporal sensory descriptions of products. Food Quality and Preference, 86.

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.

Mahieu, B., Schlich, P., Visalli, M., & Cardot, H. (2021). A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data. Food Quality and Preference, 93.

Mahieu, B., Visalli, M., & Schlich, P. (2021). Identifying drivers of liking and characterizing the ideal product thanks to Free-Comment. Manuscript in revision for Food Quality and Preference.

Visalli, M., Mahieu, B., Thomas, A., & Schlich, P. (2020). Automated sentiment analysis of Free-Comment: An indirect liking measurement? Food Quality and Preference, 82.

Visalli, M., Mahieu, B., Thomas, A., & Schlich, P. (2020). Concurrent vs. retrospective temporal data collection: Attack-evolution-finish as a simplification of Temporal Dominance of Sensations? Food Quality and Preference, 85.

Oral communications in international scientific conferences

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

Mahieu, B., Visalli, M., Thomas, A., & Schlich, P. (2020). Free-Comment methodology in consumer research: statistical analyses through the MR-Chi² framework, stability as compared to CATA and extension to temporal description and ideal profiling. In, EuroSense 2020: A Sense of Innovation.

Mahieu, B., Schlich, P., Thomas, A., & Visalli, M. (2021). Attack-Evolution-Finish: a new method allowing for temporal product description thanks to Free-Comment with consumers. In, 14th Pangborn Sensory Science Symposium.

Mahieu, B., Schlich, P., Visalli, M., & Cardot, H. (2021). A modified chi-square framework for the analysis of multiple-response data: application to Free-Comment and Check-All-That-Apply sensory data. In, AgroStat 2021.

Workshops and tutorials

Mahieu, B. (2020). Sensory characterization of home-perfumes using Free-Comment as response to open-ended questions. In, Sensometrics 2020 workshop: Applying Text Mining Methods for Sensory Evaluation Research.

Mahieu, B., & Schlich, P. (2020). Free-Comment data & analysis. In, Sensometrics tutorials at EuroSense 2020.

Posters in international scientific conferences

Mahieu, B., Visalli, M., Thomas, A., & Schlich, P. (2019). Free comment using speech recognition: an alternative to Check-All-That-Apply for sensory characterization of red wines at home. In, 13th Pangborn Sensory Science Symposium. Edinburgh, Scotland.

Mahieu, B., Visalli, M., Thomas, A., & Schlich, P. (2020). Free-Comment methodology in consumer research: statistical analyses through the MR-Chi² framework, stability as compared to CATA and extension to temporal description and ideal profiling. In, EuroSense 2020: A Sense of Innovation.

Mahieu, B., Schlich, P., Visalli, M., & Cardot, H. (2021). A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data. In, *14th Pangborn Sensory Science Symposium*.

Mahieu, B., Visalli, M., Thomas, A., & Schlich, P. (2021). Sensory evaluation with consumers revisited thanks to Free-Comment. In, 14th Pangborn Sensory Science Symposium.

Software

Benjamin Mahieu (2021). MultiResponseR: Analysis of multiple-response contingency data. R package version 1.0.0.

Publication outside the scope of this thesis

Mahieu, B., Visalli, M., Schlich, P., & Thomas, A. (2019). Eating chocolate, smelling perfume or watching video advertisement: Does it make any difference on emotional states measured at home using facial expressions? Food Quality and Preference, 77, 102-108.

Table of contents

Chapter I:

Introduction

A. Basics of sensory analysis

1. Definition

Sensory analysis is defined by the ISO norm 5492 as the "*science involved with the assessment of the organoleptic attributes of a product by the senses".* In other words, sensory analysis aims to characterize the perceptible characteristics of a product, or more generally a stimulus, using the senses of humans as the measurement instrument.

2. A bit of history

Sensory analysis is a relatively recent science that arose in the 1940-1950s to improve the sensory quality of military rations delivered to the soldiers of the US army (Jones, Peryam, & Thurstone, 1955; Peryam, Pilgrim, & Peterson, 1954). During its first applications, sensory analysis was not designated this way but rather with generic designations like "*food acceptance*" or "*food preferences*". It appears that the designation "*sensory analysis*" was first mentioned in 1961 (Depledt & Sauvageot, 2002) and lasts since. Sensory analysis met a huge rise in the 1970s by being included in R&D processes of the food-processing industry to increase the sensory quality of the developed products. Nowadays, sensory analysis is a well-established and recognized discipline, with its scientific conferences and journals, having a large community coming from various industrial and academic sectors.

3. Applications and aims

From an industrial point of view, sensory analysis is mainly used for guiding product development. It can be used to design appreciated products by the consumers, check for the sensory quality of products, investigate the impact of modifications of the formulations, compare existing products to the concurrence, etc. It can also be used to help marketing to improve the communications about the products. From an academic point of view, sensory

analysis can be used to understand feeding behaviors, establishing recommendations crossing acceptability and health or environment, studying the relation between perception and chemical compositions or biological stimulations, etc.

Due to its history, sensory analysis was initially mainly used in the food-processing industry but it is nowadays used in various sectors including cosmetic, fine fragrance, textile, home-care, pet food, sport-wear, tobacco, packaging, automobile, advertisement, and probably many others.

4. Sensory analysis in practice

From a practical point of view, sensory analysis studies are conducted following these steps:

- Stating the aims of the study
- Delimiting the product space
- Choosing the sensory method
- Establishing panelist prerequisites and recruiting them
- Planning the experimentation
- Conducting the experimentation
- Analyzing the gathered data
- Interpreting the results of the analyses and reporting them

The sensory method and the panel prerequisites are tightly linked to the aims of the study.

5. Classical sensory analysis

a. Discriminative methods

Discriminative methods aim to determine whether two or several products are different from each other. Several discriminative methods exist but they are all based on the same rationale: after the evaluations of the products, panelists are instructed to determine the products that are the same and those that are different.

These methods are usually employed when the potential differences between the products under consideration are subtle and they usually consider the products as a whole, i.e. they do not focus on any specific sensory characteristics. As discriminative methods, triangular tests, tetrad tests, and two out of five tests can be mentioned among others. Discriminative methods do not require any specific training of the panelists. The number of panelists involved in the study depends on the size of the differences between the products and on the level of statistical risk of drawing wrong conclusions considered acceptable (Rousseau, 2015).

b. Descriptive methods

Descriptive methods aim to establish the sensory profile of the products under consideration. Conventional sensory profiling is the reference descriptive method (ISO norm 13299) and consists of conducting the following actions:

- Establishing the most possible exhaustive list of descriptors that enables the scope of every single sensory characteristic of the products. From a practical point of view, this list is established thanks to existing knowledge of the products (literature, previous studies, etc.) or thanks to pre-evaluations with the panelists involved in the study. The list usually contains from 5 to 20 descriptors that can be aggregated into different sensory modalities like textures, basic tastes, aromas, etc.
- Instructing the panelists to rate the intensity of each descriptor for each product using a continuous quantitative scale. The scale might be structured or not. Products are usually presented to panelists following a monadic sequential design balanced for order and carry-over effects and with repetitions to evaluate the repeatability of the measurements.
- Establishing the sensory profile of each product based on the ratings of the panelists.
- Comparing the sensory profiles to each other using univariate and/or multivariate statistical analyses.

Sensory profiling finds its origins in *flavor profile* (Cairncross & Sjostrom, 1950), *texture profile* (Brandt, Skinner, & Coleman, 1963), *quantitative descriptive analysis* (Stone, Sidel, Oliver, Woolsey, & Singleton, 1974), and *spectrum* scales (Meilgaard, Civille, & Carr, 1991; Muñoz & Civille, 1992).

To be as objective and reliable as possible, sensory profiling must be conducted with trained panelists. The aim of the training is twofold: ensure that the panelists have a shared and suited definition of each descriptor and familiarize them with the use of the scale. In other words, the training of the panelists is the calibration of the measurement instrument. The recommended number of panelists depends on the size of the differences between the products but it usually ranges between 5 and 20 panelists (Gacula Jr & Rutenbeck, 2006; Heymann, Machado, Torri, & Robinson, 2012; Silva, Minim, Silva, & Minim, 2014; Strigler et al., 2009).

c. Hedonic methods

Hedonic methods aim to characterize the products under consideration from a hedonic point of view, i.e. to measure to which extent the products are appreciated or not. Contrary to sensory characteristics that are objective once well-defined, hedonic appreciation is always subjective since it depends on individuals' preferences. Two major types of hedonic methods can be mentioned:

- The relative methods where products are compared relatively to each other, i.e. panelists are instructed to rank the products from the least to the most appreciated without any quantification. The entire products under interest can be ranked with the ranking method. Alternatively, products can be presented and ranked by pair with the paired comparison method.
- The absolute methods where products are each rated on a *liking* scale by each panelist. Products are usually presented to panelists following a monadic sequential design balanced for order and carry-over effects.

The scale might be discrete or continuous with different levels or lengths and be labeled at regular intervals or not. However, the historical *9-points hedonic scale* (Jones et al., 1955) appears to be the most popular and commonly used.

Hedonic methods usually consider the products as a whole but they can be focused on one or several specific characteristics of the products like texture, taste, visual aspect, etc. Hedonic methods do not require any specifying training of the panelists and they are usually conducted with consumers of the target market (Stone & Sidel, 1993). The recommended number of consumers highly depends on the size of the differences between the products but it usually ranges between 60 and 100 consumers (Mammasse & Schlich, 2014).

d. Understanding preferences: linking descriptive and hedonic data

Because hedonic appreciation is one of the most important drivers of the commercial success of a product, designing appreciated products is a major problem in sensory analysis. Since hedonic appreciation is driven by the sensory characteristics of the products (Lagrange & Norback, 1987), it is necessary to link descriptive and hedonic data to understand consumers' preferences.

In the classical sensory analysis, the most popular methods to establish this linking are preference-mapping techniques (Carroll, 1972; Danzart, 2009; Greenhoff & MacFie, 1994; McEwan, 1996; Schlich & McEwan, 1992). Preference mapping techniques combine the information of two datasets: one from a trained panel that performed a descriptive method and one from an untrained panel of consumers that performed a hedonic method. Preference mapping techniques aim to provide an intuitive visual tool enabling a quick and easy diagnostic about the relation between the descriptive information and the hedonic one.

Two major approaches, differing in the point of view they adopt, can be distinguished among preference mapping techniques:

- The internal preference mapping, which focuses on the hedonic data: the product configuration is derived from liking scores and the sensory profile scores are regressed into this space.
- The external preference mapping, which focuses on the descriptive data: the product configuration is derived from sensory profile scores and the liking scores are regressed into this space.

6. Temporal sensory analysis

Sensory perception is not a static phenomenon but rather a dynamic one. In this context, several methods have been developed to measure the kinetic of the sensory perception elicited by a product intake over time. These methods can be classified into two categories: quantitative-based methods and qualitative-based methods. This section presents the most noteworthy methods of these two categories without going into detail.

a. Quantitative-based methods

The oldest methods of temporal sensory analysis were quantitative-based. Among them, the oldest is *Time-Intensity* (Lee & Pangborn, 1986) whose first applications, without being such named, were between the 1930s and the 1960s (Holway & Hurvich, 1937; Jellinek, 1964; Sjöström, 1954). Time-Intensity consists in instructing panelists to report the intensity of one sensory descriptor over time during the intake of a product using a quantitative scale. Because panelists are focused only on a single sensory descriptor at each intake, evaluating the temporal kinetic of several sensory descriptors requires several intakes. This makes Time-Intensity time-consuming and very expensive.

To overcome, to some extent, the time-consuming aspect of Time-Intensity, *Dual-Attribute Time-Intensity* (Duizer, Bloom, & Findlay, 1996), and more

recently, *Multi-Attribute Time-Intensity* (Kuesten, Bi, & Feng, 2013) were developed. The rationale of these two methods is the same as Time-Intensity except that panelists are focused on several sensory descriptors during a single intake.

To be concentrated on the intensity of the sensory descriptors continuously over time is highly demanding for the panelists. To limit this demanding aspect, *Discontinuous Time-Intensity* (Clark & Lawless, 1994) and *Progressive Profiling* (Jack, Piggott, & Paterson, 1994) were developed. The rationale of these methods is that time is discretized and panelists report the intensity of sensory descriptors at specific and predetermined times of the intake. These methods are of particular interest with relatively long-intake products such as chewing gums (Galmarini, Symoneaux, Visalli, Zamora, & Schlich, 2016).

Quantitative-based methods of temporal sensory analysis require being conducted with trained panelists for the same reasons as those mentioned for sensory profiling. In other words, panelists must be calibrated before performing the measures. The training is likely to be even more tedious than that of sensory profiling because of the temporal component. This makes quantitative-based methods of temporal sensory analysis difficult to conduct effectively.

b. Qualitative-based methods

Qualitative-based methods of temporal sensory analysis are more recent than their quantitative homologs. They arose from the assessment that quantitative-based methods are tedious and that there was a need for less complex methods to measure the kinetic of the sensory perception.

The two most popular qualitative-based methods are Temporal Dominance of Sensations (TDS) (Pineau, Cordelle, Imbert, Rogeaux, & Schlich, 2003; Pineau et al., 2009) and Temporal-Check-All-That-Apply (TCATA) (Castura, Antúnez, Giménez, & Ares, 2016). These two methods share the principle of

instructing the panelists to report their temporal sensory perception without any quantification using a presence/absence rationale where sensations (sensory descriptors) are selected by the panelists among a list of relevant ones. Originally, TDS was a quantitative-based method where the intensity of the present sensations was rated. However, facing the tediousness of quantitative ratings, only the presence/absence rationale of TDS was kept over years, letting the quantitative aspect (Schlich, 2017). TDS and TCATA differ in the number of sensations that may be cited at each time of the evaluation. TDS instructs panelists to select a single so-called "*dominant*" sensation at each time. This dominant sensation is defined as "*the sensation that catches the attention*". TCATA instructs panelists to select all sensations that "*describe*" the product at each time with the possibility of unselecting the sensations. Temporal Dominance of Sensations by Modality (Agudelo, Varela, & Fiszman, 2015), where panelists perform one TDS run for each sensory modality (texture, basic tastes, etc.), appears as a compromise between TDS and TCATA.

B. Sensory analysis with consumers

1. Motivations

Conventional sensory profiling is a very performant, reliable and robust descriptive method to determine with precision the sensory characteristics of a set of products. However, this level of accuracy has a huge counterpart: it requires trained panelists. Training a panel and further maintaining it trained over time turns out to be very expensive and time-consuming from a practical point of view. Further, the aims of some studies do not always require the high level of accuracy provided by conventional sensory profiling. This makes conventional sensory profiling not cost-effective at all in several practical situations.

Additionally, trained panelists have more to do with calibrated and objective measurement instruments than with the consumers that are, all things considered, the final users and buyers of any product. In this context, gathering

information on the consumers' sensory perception might be of paramount interest and even more relevant depending on the aims of the study. Indeed, it is likely that the consumers' sensory perception comes with a different prism than the one considered in sensory profiling.

Based on the two previous assessments, several new sensory methods have been developed over the last decades to overcome the limitations of classical descriptive sensory analysis. These methods can be classified into the following categories: descriptive methods, holistic methods, reference-based methods, hedonic-related methods, and temporal methods. This section proposes to present these methods without going into detail to position the Free-Comment method, which is the core of this thesis, among the large range of sensory methods and their corresponding particularities, rationales and aims.

2. Descriptive methods

Descriptive methods share the rationale that sensory perception results from the combination of a finite number of identifiable sensations. Each of these sensations is thus measured for each product through a sensory descriptor in an analytical way by the panelists. What makes descriptive methods different from each other is the way sensory descriptors are established and measured.

a. Intensity scales

With intensity scales, sensory descriptors are established thanks to existing knowledge of the products (literature, previous studies, etc.) or thanks to pre-evaluations. Panelists rate the sensory descriptors for each product using a quantitative scale. The products are usually presented to panelists following a monadic sequential design balanced for order and carry-over effects. The gathered data are thus the same as the ones gathered in conventional sensory profiling. The difference between conventional sensory profiling and intensity scales with consumers is that panelists are calibrated for the former while not for the latter. Using intensity scales with consumers was historically criticized (Lawless &

Heymann, 1999; Meilgaard et al., 1991; Stone & Sidel, 1993) but some more recent studies showed that this enables to provide similar product configuration, average sensory profiles and reproducibility to conventional sensory profiling (Ares, Bruzzone, & Giménez, 2011; Husson, Le Dien, & Pagès, 2001; Worch, Lê, & Punter, 2010). However, consumers' ratings show high variability and they are less consensual than calibrated panelists (Ares, Bruzzone, et al., 2011; Worch, Lê, et al., 2010). Thus, to be reliable, intensity scales with consumers require much larger panels than conventional sensory profiling to compensate for the consumers' heterogeneity. Alternatively, specific intensity scales that are easier to comprehend for not calibrated panelists such as "*labeled magnitude scales*" (Green, Shaffer, & Gilmore, 1993) may be used.

b. Free choice profiling and repertory grid

With free choice profiling (FCP) (Williams & Langron, 1984) and repertory grid (RG) (Thomson & McEwan, 1988), each panelist establishes his list of sensory descriptors and then rates them for each product using a quantitative scale. The products are usually presented to panelists following a monadic sequential design balanced for order and carry-over effects. FCP and RG differ in the way panelists establish their sensory descriptors. In FCP, they establish them based on pre-evaluations of the products while in RG they establish them based on triads of products through something close to Kelly's repertory grid (Kelly, 1955). The rationale behind FCP and RG is that panelists' sensory perception is the same but panelists differ in the way they verbalize it.

c. Flash profiling

With flash profiling (FP) (Dairou & Sieffermann, 2002), each panelist establishes his own list of sensory descriptors and then ranks the products from the least to the most intense with possible ties on each of these descriptors. The entire products under interest are thus presented at the same time to panelists. Panelists establish their sensory descriptors the same way as in FCP, i.e. based on

pre-evaluations of the product space. The rationale behind FP is that comparing products relatively to each other into a ranking task is easier than rating them in absolute.

d. Paired comparison

With paired comparison (Brard & Lê, 2016; Courcoux, Chaunier, Valle, Lourdin, & Séménou, 2005; Poirson, Petiot, & Richard, 2010), sensory descriptors are established thanks to existing knowledge of the products (literature, previous studies, etc.) or thanks to pre-evaluations. The products are presented by pairs to panelists, usually following an incomplete balanced design. For each presented pair of products, panelists determine which of the two products is the most intense regarding each sensory descriptor. The propensity of a given product of "winning a duel" regarding a given descriptor is considered as reflecting the intensity of this descriptor for this product. The rationale behind paired comparison is similar to that of FP: comparing products relatively to each other is easier than rating them in absolute.

e. Check-all-that-apply

With Check-All-That-Apply (CATA) (Adams, Williams, Lancaster, & Foley, 2007), sensory descriptors are established thanks to existing knowledge of the products (literature, previous studies, etc.) or thanks to pre-evaluations. Panelists check all sensory descriptors that apply for each product without any quantification. The products are usually presented to panelists following a monadic sequential design balanced for order and carry-over effects. The proportion of citation of a given descriptor for a given product at the panel level is considered as reflecting the intensity of this descriptor for this product. Some authors tend to confirm that this implicit assumption is indeed effective (Jaeger, Chheang, Jin, Roigard, & Ares, 2020; Vidal, Ares, Hedderley, Meyners, & Jaeger, 2018). The rationale behind CATA is that it is easier to describe a product based on a presence/absence principle rather than based on quantitative measurements.

f. Free-comment

With Free-Comment (FC) (ten Kleij & Musters, 2003), panelists describe the products with their own terms and a list of sensory descriptors is established *a posteriori* based on a dedicated pretreatment. It is then determined whether each description contains or not each descriptor resulting from the pretreatment. The products are usually presented to panelists following a monadic sequential design balanced for order and carry-over effects. The proportion of citation of a given descriptor for a given product at the panel level is considered as reflecting the intensity of this descriptor for this product. Considering that some authors tend to confirm that this implicit assumption is indeed effective for CATA (Jaeger et al., 2020; Vidal et al., 2018) and that FC data are based on the same presence/absence principle as CATA one, it is likely that this implicit assumption is indeed effective. The rationale behind FC is that freely describe a product is one of the most natural tasks that exist on the one hand, and that not using a pre-established list of sensory descriptors avoids several inherent limitations to this list on the other hand. FC is given more particular attention in section C of this chapter and will be developed throughout this manuscript, as it is the topic of this Ph.D. work.

3. Holistic methods

Holistic methods share the principle of presenting the entire products under interest at the same time to panelists and to measure similarities or dissimilarities between the products based on their overall sensory properties and without any analytical characterization as opposed to descriptive methods. Holistic methods are of particular interest when the number of products is relatively large to avoid the cognitive heaviness and resulting fatigue of descriptive methods. Further, holistic methods can catch latent information that is difficult to verbalize into analytical sensory descriptors. A free descriptive step often complements the holistic methods afterward to help to understand the product configuration, but

this is extra information. Holistic methods differ in the way distances between the products are gathered.

a. Free sorting

Free sorting originates from the field of psychology (Hulin & Katz, 1935) and it was then brought into the field of sensory analysis (Lawless, 1989; Lawless, Sheng, & Knoops, 1995). With free sorting, panelists constitute groups of products that are similar according to their own prism of perception. The groups are mutually exclusive and not subject to any restriction in terms of their size. It is usually instructed panelists to evaluate first all the products before starting to sort them. These first evaluations usually follow a design balanced for order and carry-over effects and intend to familiarize the panelists with the products. During the sorting, panelists are usually allowed to evaluate the products as many times as they need to. Panelists are usually instructed to constitute at least two groups and they are not allowed to constitute one group per product. Once products are sorted, panelists may be instructed to provide a free description of each group with few terms, the latter having been referred as *labeled sorting* (Bécue-Bertaut & Lê, 2011).

b. Projective mapping

Projective mapping (PM) (Risvik, McEwan, Colwill, Rogers, & Lyon, 1994) received several names: *placing* (Dun-Rankin, 1983), *spatial arrangement procedure* (Goldstone, 1994) in the field of psychology and *Napping*® (Pagès, 2005). Despite this diversity of designation, all these methods are based on the same principle, referred to as PM for sake of concision. With PM, panelists position the products according to their own prism of perception on a delimited rectangular area such that the distance between two products is inversely proportional to their perceived similarity. It is usually instructed panelists to evaluate first all the products before starting to position them. These first evaluations usually follow a design balanced for order and carry-over effects and

intend to familiarize the panelists with the products. During the positioning, panelists are usually allowed to evaluate the products as many times as they need to. Once products are positioned, panelists may be instructed to write down few terms to explain their mapping, which has been referred to as *ultra-flash profiling* (Perrin et al., 2008).

4. Reference-based methods

Reference-based methods share the principle of characterizing the products under interest through comparing them to reference products. The products under interest are usually presented to panelists following a monadic sequential design balanced for order and carry-over effects. Reference-based methods differ in the number of reference products they employ as well as the way products under interest are compared to reference ones.

a. Polarized sensory positioning

Polarized sensory positioning (PSP) was developed by Teillet, Schlich, Urbano, Cordelle, and Guichard (2010) following a need for a more effective sensory characterization of different waters. With PSP, panelists rate the degree of similarity of each product under interest with each reference product, the so-called "*poles*", using a continuous scale ranging from "*exactly the same*" to "*totally different*". The reference products act as latent sensory descriptors. In their original proposition, Teillet et al. (2010) used three reference products but PSP might be conducted with more or fewer references.

b. Pivot profile©

Pivot profile© (PP) (Thuillier, 2007) comes from the field of wine where free descriptions are common. With PP, panelists report, using free descriptions, the sensory characteristics that are more intense and/or less intense for each product under interest as compared to the single reference product, the so-called "*pivot*". The reference product is expected to act as an expression driver leading

PP to gather more information than Free-Comment. However, choosing an appropriate reference product is a difficult task that likely affects the ability of PP to discriminate products and to provide stable sensory characterizations of them (Brand et al., 2020).

5. Ideal-related methods

Ideal-related methods are based on the assumption that, for a given product category, consumers have in their minds a so-called "*ideal product*" that theoretically maximizes their hedonic appreciation. This assumption was first formulated by Moskowitz (1972) that suggested that consumers could evaluate the direction and the magnitude of the discrepancies between a set of actual products and their ideal product based on a set of sensory descriptors.

Ideal-related methods intend to identify as accurately as possible the product formulations that maximize consumers' hedonic appreciation. Ideal-related methods share the principle of measuring the discrepancies between the actual products under interest and the ideal one regarding a set of sensory descriptors but differ in the way the discrepancies are measured. The sensory descriptors are established thanks to existing knowledge of the products (literature, previous studies, etc.) or thanks to pre-evaluations. The actual products are usually presented to panelists following a monadic sequential design balanced for order and carry-over effects.

a. Just-about-right scales

With just-about-right (JAR) scales (e.g. (Popper, 2014)) discrepancies are measured directly and panelists rate the intensity of each sensory descriptor for each product relatively to their ideal, usually using a 5-points discrete bipolar scale ranging from a "*not enough at all*" to "*way too much*" and centered on "*just about right*". The absolute intensity of the sensory descriptors is measured neither for actual products nor for the ideal one.

b. Ideal profile method

With the ideal profile method (IPM) (van Trijp, Punter, Mickartz, & Kruithof, 2007; Worch, Lê, Punter, & Pagès, 2013), discrepancies are measured indirectly and panelists rate the intensity of each sensory descriptor for both the actual products and the ideal one using a quantitative scale as in the intensity scales method. The principle of IPM consists in characterizing the ideal product the same way as the actual products and has been successfully extended to Check-All-That-Apply (Ares, Dauber, Fernandez, Gimenez, & Varela, 2014; Ares et al., 2017; Ares, Varela, Rado, & Giménez, 2011; Bruzzone et al., 2015), projective mapping (Ares, Varela, et al., 2011) and paired comparison (Brard & Lê, 2016).

6. Temporal methods

Most of the time, temporal sensory analysis with consumers is conducted thanks to the qualitative-based methods Temporal Dominance of Sensations (TDS) or Temporal-Check-All-That-Apply (TCATA). Indeed, TDS and TCATA were originally designed for being used with trained panelists but it is now admitted that they can also be used successfully with (untrained) consumers (Ares et al., 2016; Dinnella, Masi, Zoboli, & Monteleone, 2012; Hutchings, Foster, Grigor, Bronlund, & Morgenstern, 2014; Jaeger et al., 2018; Rodrigues et al., 2016; Schlich, 2017). Rendering TDS easier to perform for consumers contributed to keep only the presence/absence rationale of the original TDS and to let its quantitative aspect aside (Schlich, 2017).

Recently, Visalli, Mahieu, Thomas, and Schlich (2020b) proposed the Attack-Evolution-Finish (AEF) method as an alternative for temporal sensory analysis with consumers. AEF instructs consumers to retrospectively select from a predefined list of descriptors the sensation they perceived at the beginning (Attack), at the middle (Evolution), and at the end (Final) of the intake of each product. The motivation for introducing AEF was to standardize and discretized *a priori* the temporal perception to avoid individual differences in terms of

response delays, mind processing of the sensory perception, and duration of the sensory perception. These individual differences are due to the continuous-time in TDS and TCATA and might noise the information. Another motivation was to render the data gathering procedure more self-explicit and easy to understand for consumers than TDS and TCATA for which a briefing phase is often performed and/or recommended (Albert, Salvador, Schlich, & Fiszman, 2012; Hutchings et al., 2014; Jaeger et al., 2017; Rodrigues et al., 2016; Thomas, Visalli, Cordelle, & Schlich, 2015). Visalli et al. (2020b) compared AEF to TDS in a study on five dark chocolates. In this study, AEF provided product discrimination and characterization very close to that of TDS.

C. Sensory analysis with consumers using Free-Comment

1. Origins and motivations

To the best of our knowledge, the first reported study using Free-Comment (FC) in the context of sensory analysis with consumers was that of ten Kleij and Musters (2003). The motivations of these authors lied in the fact that free responses "*are not often used for detailed analyses*" while they "*undoubtedly contain very rich information*" with the additional benefit of being "*stated in consumer language*". Their study confirmed that FC indeed provides rich information able to characterize a product space with a "*striking*" agreement to conventional sensory profiling.

The main practitioners' motivations for using FC rather than other descriptive methods with consumers are that FC does not rely on a pre-established list of sensory descriptors and it is based on a natural descriptive presence/absence principle. This presence/absence principle is easier and faster for consumers than ratings and rankings because it is cognitively lighter. Further, because FC does not rely on a pre-established list of sensory descriptors, it provides less biased descriptive sensory information than CATA, which is the list-based presence/absence method.

2. Benefits of Free-Comment

a. Avoiding the limitations from lists of sensory descriptors

Using a pre-established list of sensory descriptors induces several biases, thus not using such a list might be the most important benefit of FC.

Lists of sensory descriptors are likely to steer consumers in some directions and suggest to them sensory descriptors they would not have thought without the list (Coulon-Leroy, Symoneaux, Lawrence, Mehinagic, & Maitre, 2017; Kim, Hopkinson, van Hout, & Lee, 2017; Krosnick, 1999; Reja, Manfreda, Hlebec, & Vehovar, 2003; Schuman & Presser, 1979; Züll, 2016). On the contrary, FC enables the gathering of spontaneous unbiased descriptions (Lebart & Salem, 1994) that are not influenced by the practitioners and their preselection of possible applicable sensory descriptors (Foddy, 1993; Reja et al., 2003). In extreme cases of influence, the descriptive sensory information gathered by list-based methods could simply be the confirmation of practitioners' expectations (Züll, 2016). This occurs when the list of sensory descriptors is not properly established and thus it does not let the opportunity to consumers to disagree with the practitioners. Further, if the list does not enable consumers to report what they indeed perceive, the descriptive sensory information gathered is inevitably biased by the dumping effect (Campo, Ballester, Langlois, Dacremont, & Valentin, 2010; Coulon-Leroy et al., 2017; Krosnick, 1999; Varela et al., 2018). This dumping effect occurs when consumers cannot report what they perceive because it does not belong to the proposed sensory descriptors. In those situations, consumers report the sensory descriptors they judge the closest to what they perceive. Depending on the difference between the perception and the sensory descriptors of the list, the dumping effect can lead to strong misinterpretations.

To avoid the previous limitations, an "other" option might be included in the list. This additional option would aim to invite consumers to volunteer their own sensory descriptors if the ones proposed in the list do not appear relevant to

them for describing the products under interest. However, it is unlikely to conduct to the expected results as these options are generally ignored (Castura, 2009). Consumers are likely to restrict themselves to the sensory descriptors listed, even if the most appropriate sensory descriptors to describe their perception are not included in the list (Krosnick, 1999; Reja et al., 2003; Schuman & Presser, 1979; Schuman & Scott, 1987). This might result in missing some information. On the contrary, FC reduces the risk of missing some key information as the consumers are somewhat forced to volunteer their own sensory descriptors without the possibility of taking refuge in those of the list (Reja et al., 2003; Schuman & Presser, 1979).

The order in which sensory descriptors are presented in the list, as well as the size of the list, affect the attitude of consumers toward reporting their perception and thus the resulting sensory characterizations (Ares et al., 2013; Nguyen, Næs, & Varela, 2018; Varela et al., 2018). Consumers are likely to select the first reasonable and possible sensory descriptors they encounter when examining the list rather than carefully processing all possible alternatives into their minds (Krosnick, 1999). This implicitly creates an order of importance between the sensory descriptors of the list, increasing with the size of the list, which results in the first proposed sensory descriptors to be more often selected (Ares et al., 2013; Ayidiya & McClendon, 1990; Becker, 1954; Campbell & Mohr, 1950; Israel & Taylor, 1990; Kim et al., 2017; Krosnick, 1999; Krosnick & Alwin, 1987; Pineau et al., 2012). This attitude is induced by the tendency of the consumers to have a weak willingness to optimally reporting their perception (Krosnick, 1991, 1999; Krosnick & Alwin, 1987). Another bias is induced by the order in which sensory descriptors are presented: the perceptual contrast effects (Krosnick, 1999; Schwarz & Hippler, 1991). Perceptual contrast may cause a moderately applicable sensory descriptor to seem less applicable if considered

after a highly applicable one, or more applicable if considered after a high inapplicable one.

The tendency of the consumers to have a weak willingness to optimally reporting their perception might also result in biasing their thoughts in a confirmatory direction (Callegaro, Murakami, Tepman, & Henderson, 2015; Kim et al., 2017; Klayman & Ha, 1987; Koriat, Lichtenstein, & Fischhoff, 1980; Yzerbyt & Leyens, 1991). This makes consumers inclined to judge sensory descriptors applicable regardless of their indeed applicability, rather than performing the cognitive work required to evaluate this applicability. Consumers might also report safe and/or trivial perceptions, such as "cocoa" for chocolates or "alcohol" for red wines, to avoid expending the effort necessary to consider and possibly take more risky stands. In the extreme, consumers could randomly select sensory descriptors from those proposed in list-based methods (Krosnick, 1999). This confirmatory bias is reinforced by social and politeness biases: less-educated consumers with lower social status and more polite ones tend to be more biased in the confirmatory direction (Krosnick, 1999; Schuman & Presser, 1979).

b. Practical benefits

To establish properly the list of sensory descriptors, pre-evaluations are likely the best practice (Schuman & Presser, 1979), at least when the products under interest are not well known by the practitioner. These pre-evaluations aim to gather all possible sensory descriptors and dimensions present in the products under interest to compile them into a proper list of sensory descriptors. If the list of sensory descriptors is not established in this way when the products are not well known, results might be questionable (Krosnick, 1999; Schuman & Presser, 1979; Züll, 2016). Further, it is also a good practice to pre-test the list of sensory descriptors to ensure that it is relevant (Krosnick, 1999). Thus, establishing and pre-testing properly a list of sensory descriptors can be time-consuming. In this context, from the practitioners' point of view and assuming the products under

interest are not well known, the main benefit of open questions is that they are easier and quicker to set up as they do not require extensive preparations. Further, since FC is a natural and spontaneous task, it does not require deep explanations to be understood by consumers. This makes FC a flexible and relevant method for less controlled testing conditions such as home-used tests with the benefit of producing data that can be aggregated across different studies.

From the consumers' point of view, FC might be a more motivating format, especially when the list of sensory descriptors would have been oversized to them (Züll, 2016). Similarly, FC is less demanding than list-based methods because consumers answer spontaneously without the need of mind processing every possible sensory descriptor of the list, which is less time-consuming and less cognitively heavy (Lebart & Salem, 1994). Moreover, FC puts the consumers in a climate of trust and confidence, which favor communication and thus enhance their willingness to report optimally their perception (Bradburn & Sudman, 1979; Lebart & Salem, 1994; Sudman & Bradburn, 1974).

3. Limitations of Free-Comment

The main limitation of FC is that it requires a relatively extensive pretreatment to establish an *a posteriori* list of sensory descriptors. Since it exists several ways of conveying the same descriptive sensory information with possible typing errors, this pretreatment is time-consuming and cumbersome relatively to list-based methods that do not require such pretreatment (Hanaei, Cuvelier, & Sieffermann, 2015; Payne, 1980; Reja et al., 2003; Sheatsley, 1983; Symoneaux, Galmarini, & Mehinagic, 2012; ten Kleij & Musters, 2003).

Another limitation of FC is that some consumers, generally the less educated ones, might encounter some difficulties to verbalize properly their perception (Krosnick, 1999; Reja et al., 2003). This might result in some broad and general sensory descriptors in the *a posteriori* list that bring only overall and imprecise descriptive sensory information (Schuman & Presser, 1979). Further,

some consumers might provide only hedonic information. On the contrary, when *a priori* lists of sensory descriptors are properly established, they produce only informative descriptive sensory information since practitioners do not include uninformative sensory descriptors or hedonic ones in the lists (Reja et al., 2003; Schuman & Presser, 1979), at least when they are solely interested in sensory descriptive information. Further, using a pre-established list of sensory descriptors enables to provide the consumers a definition of each of these descriptors to minimize the risk of misinterpretations of the products' characterizations. While this can be tedious and does not entirely erase any risk because of individual interpretations of the definitions, this offers an opportunity to render list-based methods less subject to misinterpretations than FC.

FC instructions require being very precise in the way they are stated as well as the most possible focused on a single aspect (Reja et al., 2003; Symoneaux et al., 2012; Züll, 2016). If they are not, the descriptions might be uninformative regarding the aspect investigated. On the contrary, list-based methods do not necessitate such preciseness in their instructions since the list of sensory descriptors guides consumers on the aspect investigated. Further, list-based methods enable to render relevant and applicable sensory descriptors considered as trivial and/or obvious by consumers while FC might miss this information due to consumers not mentioning these sensory descriptors (Lebart & Salem, 1994).

4. Popularity of Free-Comment

The other descriptive methods of sensory analysis with consumers but FC and CATA are based on rating or ranking products under interest regarding several sensory descriptors. Since this is relatively cognitively heavy and difficult for consumers and further that rating is sometimes criticized, using a presence/absence principle appears the best practice for sensory characterization with consumers. In this context, the popularity of FC is only compared to that of CATA:

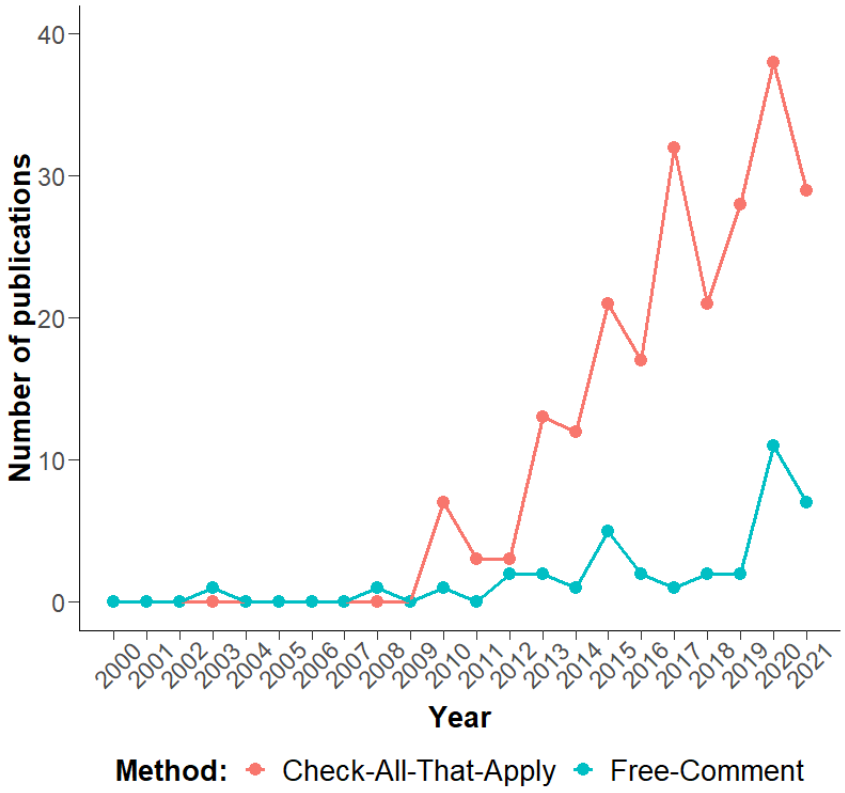


Figure 1: Number of publications related to Free-Comment and Check-All-That-Apply since 2000

Figure 1 shows the number of publications related to FC and CATA since 2000 as returned by Scopus and restricted to the following journals: Food Quality and Preference, Food Research International, Journal of Sensory Studies and Foods. For FC, since "Free-Comment" is not a consensual denomination, the following keywords were used: "free comment", "open comment", "comment analysis", "text analysis", "open ended questions" and "free text comment". For CATA, the following keywords were used: "CATA", "Check All That Apply" and "Choose All That Apply". The research was performed among article titles, abstracts and keywords. Note that no *a posteriori* manual filtering was performed meaning that the number of publications might be overestimated, especially for FC, as the research equation was quite extensive.

Figure 1 confirms that, despite its many benefits, FC is relatively unpopular as compared to CATA. This assessment suggests that the benefits of FC have less

weight than its limitations to practitioners, the heaviest limitation likely being the need for FC data to be pretreated before them being analyzable. The lack of many reported applications of FC and documentation of its performances likely reinforces the obstacles to its use and justifies the work exposed in this manuscript.

5. Data gathering

To the best of our knowledge, four different nuances of Free-Comment (FC) with consumers were reported in the literature. These nuances differ about the degree of "freedom" they let to consumers but whatever the nuance, the essence of FC, i.e. letting consumers describe the products with their own terms, is preserved. The first nuance does not impose any restriction regarding the descriptions that can be provided (ten Kleij & Musters, 2003): consumers can describe any sensory modality (visual aspect, flavor, etc.) with as many terms they wish without any imposed format of description. The second nuance imposes on consumers to focus their descriptions on a single sensory modality (Hanaei et al., 2015). The third nuance imposes on consumers to provide hedonic-oriented descriptions by categorizing their descriptions into a "*like*" category and a "*dislike*" one (Lahne, Trubek, & Pelchat, 2014; Symoneaux et al., 2012) or by constraining their descriptions to a Just-About-Right scale syntax (Luc, Lê, & Philippe, 2020). The fourth nuance imposes on consumers to provide a limited number of terms (Ares, Giménez, Barreiro, & Gámbaro, 2010).

6. Pretreatment

The pretreatment of FC data aims to clean and standardize the descriptions into a list of sensory descriptors for the panel. Five key steps are shared by the reported approaches to achieve this aim (Ares et al., 2010; Hanaei et al., 2015; Lahne et al., 2014; Symoneaux et al., 2012; ten Kleij & Musters, 2003):

- To check for and to correct typing and spelling errors.

- To remove uninformative information (e.g. punctuations, stop words, etc.).
- To lemmatize the corpus, i.e. to turn every occurrence of a term into its canonical form i.e. its usual form findable in the dictionary.
- To group synonym terms and terms that convey similar descriptive sensory information into a single term.
- To apply a threshold of citations for a term to be considered in statistical analyses to avoid unreliable sparse characterizations. Some authors considered an overall threshold of citations that is independent of the repartition of the citations over the products. For these overall thresholds, two strategies were reported: using an absolute threshold (ten Kleij & Musters, 2003) or using a relative percentage of consumers as a threshold (Ares et al., 2010; Hanaei et al., 2015). Other authors considered a repartition-dependent threshold (Lahne et al., 2014; Symoneaux et al., 2012). This latter strategy consists in considering a term in the statistical analyses if it was mentioned by at least a certain percentage of consumers for at least one same product.

Besides these shared steps, some authors took into account negations (e.g. *not*, *not very*, etc.) (Ares et al., 2010; Symoneaux et al., 2012; ten Kleij & Musters, 2003) and quantifiers (e.g. *very*, *a little*, etc.) (ten Kleij & Musters, 2003) while it seems some others did not (Hanaei et al., 2015; Lahne et al., 2014).

7. Statistical analyses

The reported studies involving the use of FC with consumers summarized the pretreated FC descriptions into a contingency table crossing the products with the sensory descriptors from the list established *a posteriori* (Ares et al., 2010; Hanaei et al., 2015; Lahne et al., 2014; Symoneaux et al., 2012). In this contingency table, each cell contains the number of times the sensory descriptor of the corresponding column was cited for the product of the corresponding row

at the panel level. The contingency table is then submitted to a Correspondence Analysis (CA) (Benzécri, 1973) to investigate the structure of the dependence between products and sensory descriptors. CA enables to depict the structure of the dependence between products and sensory descriptors according to a chi-square criterion by decomposing the dependence into orthogonal ranked axes of maximal and decreasing dependence. Usually, the first two axes are retained for interpretation and used to map the product and the sensory descriptors into an easy to comprehend bi-dimensional space. On this map, the closer two products are, the more similar their sensory characterization. The position of each product relatively to the sensory descriptors enables investigating what makes it different or similar to the other products.

While some authors rushed on investigating the structure of the dependence (i.e. performing CA) without verifying if the dependence they investigate is large enough to consider it worthy of investigation (i.e. significant) (Ares et al., 2010; Hanaei et al., 2015; ten Kleij & Musters, 2003), other authors verified it (Lahne et al., 2014; Symoneaux et al., 2012). This verification was performed based on a "*global chi-square test*" complemented by "*chi-square tests per cell*" to identify the cells having an observed count significantly different from its expected count under independence.

8. Performances

Every reported study that used FC with consumers demonstrates its ability to differentiate and characterize a set of products (Ares et al., 2010; Hanaei et al., 2015; Lahne et al., 2014; Symoneaux et al., 2012; ten Kleij & Musters, 2003). FC was shown to be able to provide similar product configuration and product characterizations to conventional sensory profiling (Ares et al., 2010; Symoneaux et al., 2012; ten Kleij & Musters, 2003). FC further enables to capture rich and sensible characterizations of the products that are relevant to consumers and in

their own language (Ares et al., 2010; Hanaei et al., 2015; Lahne et al., 2014; Symoneaux et al., 2012; ten Kleij & Musters, 2003).

Results provided by FC appear quite reproducible, at least regarding the main sensory dimensions. Indeed, in Lahne et al. (2014) consumers characterized the same products twice with different levels of information and the resulting product configurations and characterizations depicted by CA were highly similar in both contexts. In Hanaei et al. (2015), the authors added a blind duplicate to their products under interest and these duplicates were each other closest products on the CA map suggesting they had highly similar characterizations. Further, it appears that the main sensory descriptors generated spontaneously by the consumers are stable across studies if the product spaces investigated are similar (Hanaei et al., 2015).

D. Aims and structure of this manuscript

This thesis aims to put Free-Comment (FC) in the spotlight for sensory analysis with consumers. This is motivated by the several benefits of FC presented in chapter I that are not exploited in depth because its performances are not well documented and its analyses and range of application remain limited.

Chapter II first presents the FC data gathering procedure that is proposed to gather the most possible information on the products under interest. Second, it presents a semi-automatized procedure to perform the pretreatment of FC data. For this second point, particular attention is given to offer a standardized pretreatment the fastest and objective as possible while minimizing the loss of information and richness of the FC descriptions.

Chapter III presents the statistical analysis proposed to be applied to the pretreated FC data. Because the pretreated FC data have the same structure as the Check-All-That-Apply (CATA) data, the proposed analyses are also relevant for analyzing CATA data. The first section of chapter III proposes to determine and

account for the dimensionality of the dependence between products and sensory descriptors in the analysis, i.e. the number of significant Correspondence Analysis (CA) axes. Besides, it proposes to compute confidence ellipses for the products' locations in the sensory space depicted by the CA. The second section of Chapter III introduces a multiple-response chi-square framework and proposes to rely on it instead of the usual chi-square framework for the analysis of FC data. The new framework considers an evaluation (vector of citations for one product by one consumer) as being the experimental unit. This latter framework is more suited to FC data than the usual chi-square one because it is not subject to the same limitations when the products under interest elicit different rates of citations.

Chapter IV proposes to compare the performances of FC to those of CATA, which is more popular and whose performances are more documented. The first section of chapter IV compares FC and CATA in terms of product discrimination and characterization. For this comparison, two groups of consumers evaluated four red wines with a FC or a CATA protocol depending on the group they belonged to. The second section of chapter IV compares the FC and CATA in terms of the stability of the descriptive sensory information they provide. For this comparison, the previous data on the red wines were used together with the data from another study on four milk chocolates that also included a FC group and a CATA group.

Chapter V proposes two new sensory methods to extend FC to other typical situations of sensory evaluation, namely temporal sensory analysis, drivers of liking identification and ideal product characterization. The first section of chapter V tackles temporal sensory analysis by proposing the Free-Comment Attack-Evolution-Finish (FC-AEF) method. With FC-AEF, the evaluation of each product is split into three periods, the beginning (Attack), the middle (Evolution), and the end (Finish), and consumers are instructed to report their perception retrospectively using a FC description for each of these three periods. An

application of FC-AEF on five dark chocolates is presented. The second section of chapter V tackles drivers of liking identification and ideal product characterization by proposing the Ideal-Free-Comment (IFC) method paired with liking scoring. Three types of data are gathered in this method: the FC descriptions of the products under interest, the liking scores of the products under interest and the FC descriptions of the ideal product. An application of IFC paired with liking scoring in a large study involving 483 consumers evaluating from one to fourteen cooked hams from a list of 30 hams representative of the French market is presented.

Finally, chapter VI discusses the propositions and the results of these works and suggests directions for future works while chapter VII gives an overall conclusion of this thesis.

Chapter II:
Gathering and pretreatment of Free-Comment data

A. Gathering of Free-Comment data

All Free-Comment (FC) data except for temporal data have been acquired in the same way during this thesis. Consumers described the products with their own terms, as it is the essence of the FC method. No restriction was imposed on them regarding the number of terms they could use, the nature of the terms they could use, and the form of the descriptions they could provide. This decision was taken not to alter the "Free" aspect of "Free-Comment". Consumers had to provide one separate FC description for each sensory modality of the products under investigation. This decision was taken to increase the precision of the instructions (Symoneaux et al., 2012; Züll, 2016) and raising awareness of the consumers about all the characteristics of the products and thus to decrease as much as possible the probability of missing some descriptive sensory information about the products. The instructions were stated as: "*Describe the sensory modality of this product*" with "*sensory modality*" and "*product*" being replaced by the investigated sensory modalities and the type of products under interest. Hedonic-oriented FC (Lahne et al., 2014; Luc et al., 2020; Symoneaux et al., 2012) was not considered in this thesis because only "pure" and non-oriented descriptive sensory characterization of the products was under interest.

For the data gathering involving a temporal component, the previous division of the sensory perception into different sensory modalities was not performed as it was already divided into temporal periods. This decision was taken to avoid the task from being too difficult and cognitively heavy to the consumers. The instructions were stated as: "*What sensations did you perceive during the tasting (textures, flavors, aromas, etc.) in chronological order?*". The examples of sensory modalities given in brackets aimed to play the same role as the division of the sensory perception in the non-temporal FC data gathering procedures.

B. Pretreatment of Free-Comment data

Some limitations regarding the reported approaches for pretreating Free-Comment (FC) data presented in chapter I can be mentioned. Most of the time, the pretreatment of FC data was reported to be manually conducted and thus time-consuming (Ares et al., 2010; Hanaei et al., 2015; Symoneaux et al., 2012; ten Kleij & Musters, 2003). The fact that negations seem not to be taken into account systematically can lead to huge misinterpretations since for example *strong* and *not strong* is "*a big difference!*" (ten Kleij & Musters, 2003). Grouping the terms with similar meanings into a single term adds subjectivity in the procedure because a single term is selected more or less arbitrarily to represent all its synonyms. Further, it renders the grouping procedure unclear and it discards nuances of the terms provided by the consumers, which results in losing a part of the richness of the FC method. Grouping the terms with similar meanings only based on semantic considerations can be quite subjective and further time-consuming due to the need for considering every possible grouping and due to the resulting arbitration. Finally, using an overall threshold of citations for a term to be considered in statistical analyses is suboptimal since it does not guarantee any consensus from the consumers while repartition-dependent thresholds (a certain percentage of consumers for at least one same product) do to some extent.

To remedy the limitations mentioned above, a new and original pretreatment procedure was developed and is described thereafter. All the FC datasets of this thesis were pretreated with this procedure. The procedure was entirely performed using the R software (R Core Team, 2020) and using the lexicon from the IRaMuTeQ© software (Ratinaud, 2014) for lemmatization (turning terms to their canonical form) and part-of-speech tagging (identification of the grammatical class of terms). An extract of this lexicon is depicted in Table 1:

Term	Lemma	Grammatical class
…	…	…
bitter	bitter	adj
bitterer	bitter	adj
bitterest	bitter	adj
…	…	…
fruitier	fruity	adj
fruitiest	fruity	adj
fruity	fruity	adj
…	…	…
saltier	salty	adj
saltiest	salty	adj
salty	salty	adj
…	…	…

Table 1: Extract of the lexicon used for the pretreatment of Free-Comment data

Depending on the dataset, the procedure was performed by sensory modality or with aggregated periods. Figure 2 summarizes the proposed semi-automatized procedure. On this figure, "Automatized" refers to no manual intervention in the R code while "Manual intervention" refers to the opposite.

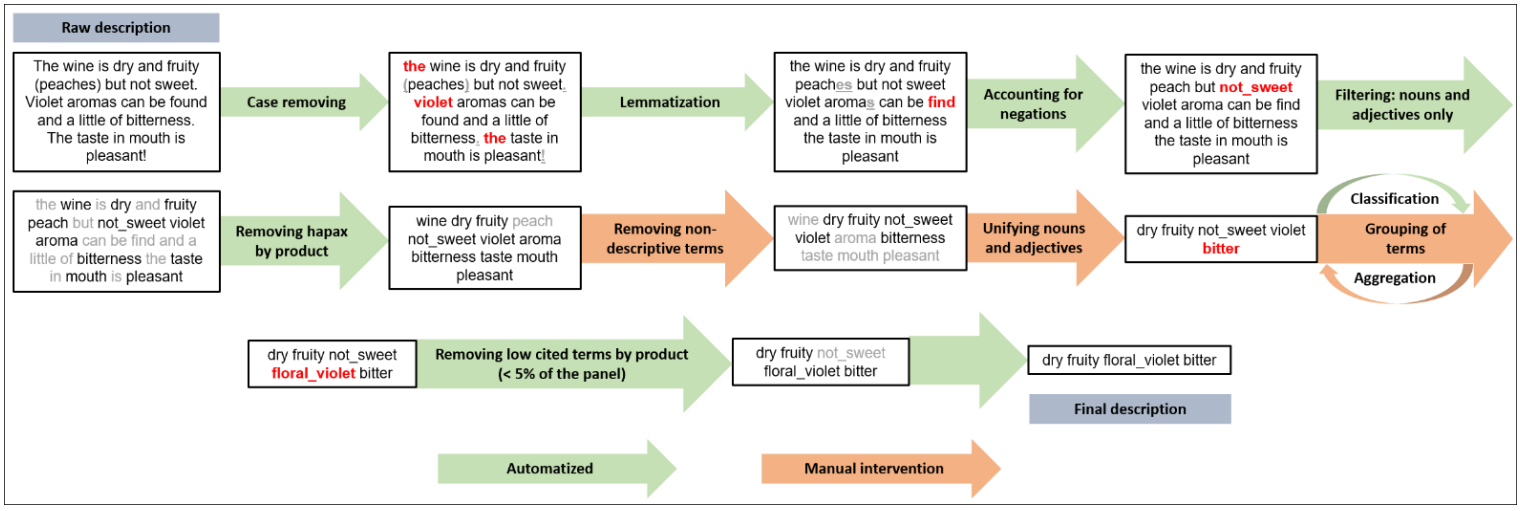


Figure 2: Example of pretreatment of Free-Comment data on a fictive description

The proposed procedure consists in successively performing the following actions on the descriptions:

- Case removing: all characters but letters are removed and letters are all turned to lower case.
- Lemmatization: all terms are turned to their canonical form, i.e. their usual form findable in the dictionary, using the lexicon from the IRaMuTeQ© software.
- Accounting for negations: all negations are linked to their associated sensory descriptive term.
- Filtering: all terms but nouns and adjectives are removed using the lexicon from the IRaMuTeQ© software.
- Removing hapax by product: terms not mentioned at least two times for at least one product at the panel level are removed.
- Removing non-descriptive terms: terms not conveying descriptive sensory information are removed. In case of ambiguity about the meaning of a term, the meaning of this term is deducted from the raw descriptions.
- Unifying nouns and adjectives having the same grammatical root
- Grouping of terms conveying similar descriptive sensory information, i.e. related to the same sensory dimension, by alternating classification and aggregation (this step is detailed after).
- Removing low cited terms by product: descriptors not mentioned at least by 5% of the panel for at least one product are removed to avoid unreliable sparse characterizations.

For the grouping of terms conveying similar descriptive sensory information, the terms are first classified based on their profile of citations i.e. their repartition of citations over the products. The method of classification used is the one proposed by Greenacre (1988) which is an ascendant hierarchical classification based on the chi-square distance and a weighted Ward merging criterion. At each step of the hierarchical tree building, two terms are merged such

as to keep the chi-square statistic of the descriptor by product contingency table as high as possible. The final classification of the terms results from the collapsed contingency table having the most significant chi-square statistic i.e. the lowest p-value. Once the terms are classified, terms conveying similar descriptive sensory information within each class are aggregated into a latent term containing all its constituting terms displayed. No aggregation of terms is performed between classes. These aggregations are performed manually to ensure they are consistent with sensoriality and semantic. Indeed, terms having similar profiles do not necessarily convey the same descriptive sensory information. Some of them simply applied to the same products and in similar proportions. Once all aggregations are performed, the classification/aggregation procedure is repeated until no more aggregation could be performed consistently with sensoriality and semantic.

A final list of sensory descriptors shared by the panel is then established as the terms and latent terms resulting from the pretreatment. Finally, the pretreated descriptions are encoded in a presence/absence (1/0) matrix where one row corresponds to a pair of consumer and product and one column corresponds to a sensory descriptor. An example of pretreated FC data is depicted in Table 2.

Consumer	Product	D_1	D_2	D_3	…	D_N_D
C_1	P_1	1	0	0	…	0
C_1	P_2	0	1	0	…	1
C_1	P_3	0	1	0	…	0
C_1	P_4	1	0	1	…	1
…	…	…	…	…	…	…
C_N_C	P_1	0	1	0	…	0
C_N_C	P_2	1	0	0	…	1
…	…	…	…	…	…	…
C_N_C	P_N_P	0	0	1	…	1

Table 2: Example of pretreated Free-Comment data

The pretreatment is performed using computer software so that some steps are entirely automatized which contributes to render the pretreatment as fast and standardized as possible. The aim of the classification is twofold. First, it facilitates the task to practitioners by making some "propositions" of aggregations. Second, it contributes to standardize the pretreatment procedure by limiting subjective aggregations of terms. Aggregating terms into latent terms containing all their constituting terms displayed also shows two aims. First, it avoids discarding shades of different terms conveying similar descriptive sensory information, as these shades are part of the richness of the FC method. Second, it clarifies the groupings of terms performed by practitioners and contributes to standardize the pretreatment procedure by avoiding the arbitrary choices of one term to represent several ones.

Chapter III:

Statistical analyses of Free-Comment data

A. Context and contents

Some limitations regarding the statistical analyses of pretreated Free-Comment (FC) data can be mentioned. The fact that dependence between products and sensory descriptors is not systematically tested for significance can lead to a strong over-interpretation. Indeed, Correspondence Analysis (CA) operates on proportions and not on counts and it consequently investigates the structure of the deviations from independence rather than the deviations themselves, i.e. it does not depend on the sample size. Thus, knowing whether one investigates significant deviations from independence is crucial (Saporta, 2006). Even if the chi-square test is significant, it only means that at least the first CA axis captures a significant dependence (Camiz & Gomes, 2013; Malinvaud, 1964; Saporta, 2006). In other words, the dependence captured by subsequent axes should also be tested and further analysis ideally restricted to significant axes only. CA of the contingency table crossing products and sensory descriptors provides an average product configuration but does not enable investigating the stability of this configuration and thus investigating the significance of pairwise discrimination of products. Finally, the usual chi-square test and the usual CA that directly comes from it are not well suited for the statistical analysis of FC. Indeed, these approaches consider the citation of one descriptor by one consumer for one product as an experimental unit, which does not fit well to FC data because some citations come from the same evaluation and they might be correlated. This chapter proposes to remedy these limitations. Note that because pretreated FC data are of the same nature as Check-All-That-Apply (CATA) data, these limitations also apply to the usual analysis of CATA, which makes the proposition of this chapter relevant to analyze CATA data too.

Section B proposes an integrated set of analyses to account for the dimensionality of the dependence between products and sensory descriptors, i.e. the number of significant Correspondence Analysis (CA) axes. A sequential

procedure for testing the dependence captured by each CA axis is proposed. It is then proposed to restrict the analysis to the significant axes. Accordingly, confidence ellipses of products' location in the CA space are proposed to be computed thanks to a total bootstrap procedure in which Procrustes rotations are performed within the significant subspace. Further, tests per cell to determine significant associations between products and sensory descriptors are proposed to be performed using Fisher's exact tests applied on the contingency table derived from the significant axes. This derived contingency table is obtained thanks to the *reconstitution formula* of CA. This integrated analysis and its benefits are demonstrated based on CATA data in Figures 1 and 2 of section B.

Section C introduces the multiple-response chi-square framework and proposes to rely on it instead of the usual chi-square framework for the analysis of FC data and CATA data. The tools developed in section B for the usual chi-square framework are generalized in section C to the multiple-response framework, i.e. a multiple-response CA with a test of dependence of its axes and a multiple-response hypergeometric test for the tests per cell. It is thus possible to account for the dimensionality of the dependence between products and sensory descriptors as proposed in section B in this new framework. The difference between the two frameworks lies in the experimental unit considered by each of them. Unlike the usual chi-square framework, the multiple-response chi-square framework considers an experimental unit as being an evaluation i.e. a vector of citations for one product by one consumer. This difference of point of view leads the expected counts under the null hypothesis of no association between products and descriptors to differ between the two frameworks. The multiple-response chi-square framework's point of view is well suited to the nature of FC and CATA data, while the point of view of the usual one is not. Thus, the expected counts are valid with the multiple-response chi-square framework's point of view, while they are not with the usual chi-square framework. From a practical point of view, this

results in the usual chi-square framework providing inconsistent and counterintuitive outputs when the products elicit different citation rates (all sensory descriptors combined). The multiple-response chi-square framework remedy this limitation as demonstrated based on CATA data in Figures 1, 2 and 3 of section C.

The analyses introduced in this chapter have been implemented into the MultiResponseR R-package, which is presented in Appendix and freely available at: https://github.com/MahieuB/MultiResponseR.

B. Accounting for the dimensionality of the dependence between products and sensory descriptors in analyses of Free-Comment data

Article published in Food Quality and Preference:

Accounting for the dimensionality of the dependence in analyses of contingency tables obtained with Check-All-That-Apply and Free-Comment

Benjamin Mahieu*, Michel Visalli, Pascal Schlich

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

Reference:

Mahieu, B., Visalli, M., & Schlich, P. (2020). 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.

Food Quality and Preference 83 (2020) 103924

Contents lists available at ScienceDirect

Food Quality and Preference

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

Accounting for the dimensionality of the dependence in analyses of contingency tables obtained with Check-All-That-Apply and Free-Comment

Benjamin Mahieu*, Michel Visalli, Pascal Schlich

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

ARTICLE INFO

Keywords:
Correspondence analysis
Dimensionality test
Monte-Carlo test
Confidence ellipses
Chi-square per cell

ABSTRACT

Check-All-That-Apply (CATA) and Free-Comment (FC) provide a so-called contingency table containing citation counts of words or descriptors (columns) by products (rows). This table is most often analysed using correspondence analysis (CA). CA aims at decomposing dependence between products and descriptors into axes of maximal and decreasing dependencies, which is reasonable if the dependence has been previously established by a chi-square test. However, the p-value of this test is not valid when the observations are not independent or when the contingency table contains too many low expected citation rates. In addition, rejecting independence with a chi-square test only means that at least the first CA axis captures some dependence. This paper presents a test to determine the number of axes of the CA that capture significant dependence and proposes a Monte-Carlo approach to compute valid p-values for this test. The variability in the products' coordinates in the CA space is often evaluated by means of a total bootstrap procedure. The paper proposes to rely on this test to determine the number of axes to consider for the Procrustes rotations of such a procedure. Finally, to investigate which words are cited more often for each product, the paper proposes performing Fisher's exact tests per cell on the derived contingency table obtained by reversing the CA computations on the axes capturing significant dependence. The benefits of accounting for the dimensionality of the dependence in the analyses are demonstrated on real CATA data.

1. Introduction

In recent years, new consumer-oriented methods have emerged to overcome the limitations of sensory descriptive analysis (Valentin, Chollet, Lelièvre, & Abdi, 2012; Varela & Ares, 2012), including word citation occurrence-based methods, which aim to collect 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. The most commonly used word citation occurrence-based methods are Check-All-That-Apply (CATA) (Adams, Williams, Lancaster, & Foley, 2007) and Free-Comment (FC) as response to open-ended questions (ten Kleij & Musters, 2003). Ultra-flash profiling (UFP) (Perrin & Pagès, 2009) and labelled sorting (Abdi & Valentin, 2007) could also be seen as word citation occurrence-based methods, but the word-based descriptive data are not the main output when using these two methods.

Data collected from a CATA or FC task are stored in a so-called contingency table containing citation counts of words or descriptors (columns) by products (rows). Each cell of the contingency table contains the number of times a product was described by a word. The first step to study such a dataset is to test for overall differences between products. In the context of contingency tables collected using FC, this is usually performed using a chi-square test (Galmarini, Symoneaux, Chollet, & Zamora, 2013; Lahne, Trubek, & Pelchat, 2014; Lawrence et al., 2013; Symoneaux, Galmarini, & Mehinagic, 2012). However, computing the p-value of the chi-square test using the chi-square distribution is valid only if the following conditions are met: (i) the observations are independent, (ii) no expected cell count is less than five in the contingency table (Agresti, 2007) and (iii) the contingency table is not sparse (Renter, Higgins, & Sargeant, 2000). In the context of contingency tables obtained using CATA or FC, these conditions are rarely met, especially the first condition, as all subjects evaluate all the products by assessing all the words. In the context of contingency tables collected using CATA, to address the issue of the non-validity of the chi-square distribution, Meyners, Castura, and Carr (2013) proposed to test for overall differences between products using a Monte-Carlo test based on combination of Cochran's Q statistics. In both contexts, if overall difference between products is not established, pursuing further analyses is not recommended. When overall difference between products is established, then a correspondence analysis (CA) (Benzécri, 1973) can

* Corresponding author.
E-mail address: benjamin.mahieu@inrae.fr (B. Mahieu).

https://doi.org/10.1016/j.foodqual.2020.103924
Received 22 November 2019; Received in revised form 27 February 2020; Accepted 28 February 2020
Available online 29 February 2020
0950-3293/

be performed to visualise the association between products and words on a factorial map that decomposes the dependence between products and words into axes of maximal and decreasing dependencies. Furthermore, it is common to represent the variability in the products' coordinates in the CA space using confidence ellipses on the CA map. Confidence ellipses can be constructed in two ways: parametric bootstrap using a multinomial distribution (Antúnez, Ares, Giménez, & Jaeger, 2016; Oppermann, de Graaf, Scholten, Stieger, & Piqueras-Fiszman, 2017; Ringrose, 2012) or total bootstrap based on resampled subjects (Alcaire et al., 2017; Cadoret & Husson, 2013; Vidal, Ares, Hedderley, Meyners, & Jaeger, 2018). There are, to the best of our knowledge, two approaches to interpret relations between products and words in an objective manner. The first approach consists of computing the chi-square per cell on the contingency table (Symoneaux et al., 2012) to list words significantly more or less cited for each product; these words contribute the most to the global chi-square statistic. The second approach consists of performing the Multidimensional Alignment (MDA) on CA coordinates to interpret the cosine of the angle between product vectors and word vectors in the full CA space (Carr, Dzuroska, Taylor, Lanza, & Pansini, 2009; Meyners et al., 2013).

When overall difference between products is established, it only means that at least the first axis of the CA captures a significant dependence. From that result, there is a need to know how many other axes capture a significant dependence. Moreover, all computations performed with the analyses presented above are performed without considering how many axes capture a sufficient dependence to be considered significant. Thus, these methods do not take into account the dimensionality of the dependence and potentially add noise or miss important information needed for the interpretation.

The present paper proposes an approach that considers the dimensionality of the dependence when analysing CATA or FC data. The first section introduces a test of dimensionality based on chi-square statistic and on a Monte-Carlo approach to compute valid p-values. Chi-square statistic was chosen over the alternative Monte-Carlo test proposed by Meyners et al. (2013) because this latter is based on combination of Cochran's Q statistics that are not related to CA. The paper then explains how to take into account the information provided by the test when investigating the variability in the products' coordinates in the CA space and the relations between products and words. In the second section of this paper, the results obtained with this new approach are compared to those provided by the traditional analyses. In the last part, the benefits and limitations of both approaches are discussed. Finally, a global conclusion is given.

2. Material and methods

2.1. Testing dependence captured by the CA axes

Because CA and chi-square statistic belong to the same rationale, they are tightly related to each other. The tight relation between CA and chi square statistic gives interesting properties that enable testing the dependence captured by the CA axes. For this reason, the subsequently proposed test relies on chi-square statistic and not the test based on combination of Cochran's Q statistics proposed by Meyners et al. (2013). Further, contrarily to the Cochran's Q test that tests for equality of citation proportions across products for a given word, the chi-square test tests for independence between products and words and thus takes into account the total numbers of citations of the products (their margins).

The chi-square statistic of a contingency table is linked to the eigenvalues of the CA performed on this contingency table by the following equation:

$$\chi^2 = N \times \sum_i \lambda_i$$

where χ^2 is the chi-square statistic of the contingency table, N is the sum of all the cells of the contingency table, and λ_i is the i-th eigenvalue of the CA.

The sum of the eigenvalues of the CA can be seen as the effect size or the absolute intensity of the dependence between rows and columns. It is equal to the chi-square statistic divided by N and is thus based only on the observed and expected probabilities of being in each cell of the contingency table. Contrary to the chi-square statistic, it is independent of the sample size. Based on the above equation, it is possible to test for the dependence of each CA axis with a stepwise procedure (Camiz & Gomes, 2013). The idea is to test, at each step, whether removing the dependence captured by the axes of all the previous steps still results in rejecting independence in the sense of the chi-square test, i.e., if there is still enough dependence to be considered significant.

Suppose that we have a contingency table X of size n × p. The rank of X is equal to the minimum of (n-1) and (p-1) or less if there is a singularity. Let us denote this rank D. Let k vary from 1 to D until independence is not rejected for an axis. The principle of the stepwise procedure is as follows:

(i) At the k-th step, compute the following statistic: $Q_k = N \times \sum_{i=k}^{D} \lambda_i$
(ii) Compare this statistic to the quantiles of a chi-square distribution with (n-k)(p-k) degrees of freedom to obtain a p-value
(iii) If this p-value is less than the predetermined α risk, then set k = k + 1.

Running this procedure until independence is not rejected provides the number of CA axes that capture some significant dependence and thus the dimensionality of the data in the sense of dependence. The statistic computed at step k = 1 is equal to the statistic of the chi-square test. At step k ($1 \leq k \leq D$), the test is conceptually equivalent to perform a chi-square test on the derived contingency table represented only by the k-th to the D-th CA axes.

In practice, as stated in introduction (Section 1), computing the p-value of the chi-square test using the chi-square distribution is not valid in the context of contingency tables collected using CATA and FC. To overcome this limitation, a Monte-Carlo approach (Adery, 1968) is proposed. In such an approach, a large number of datasets are simulated under the null hypothesis investigated and then the statistic of interest is computed for each simulated dataset. These computations enable the user to obtain an empirical distribution under the null hypothesis with no probabilistic assumption. The statistic of interest computed on the real dataset is then compared to those of the simulated distribution under the null hypothesis, and the p-value is the proportion of the simulated statistics more extreme than or equal to the observed one. Here, the null hypothesis is independence between products and words on the k-th axis and the statistic of interest is Q_k.

The simulated data under the null hypothesis must be consistent with the nature of the data. In our case, the contingency table is obtained by summing the number of citations of each word for each product across the subjects. Simulating data by considering only the information provided by the observed contingency table, using, for example, Patefield's algorithm (Patefield, 1981), omits the subjects' individual information and thus is not appropriate. To overcome this limitation, independence can be simulated by randomly reallocating each word citation to a product by subject. However, this approach is problematic because it does not take into account the semantic nature of the words, so it could lead to unrealistic individual simulated data. For example, if a subject used the words "hard" and "soft" to describe a set of products, one can hope that both of these words were not used to describe the same product, but that could happen after random reallocation. For these reasons, this approach is also not appropriate. A more appropriate alternative to simulate consistent data consists of considering whole descriptions instead of words. Here, a description refers to the set of words used by one subject to describe one product. As these descriptions are indeed observed, they are realistic from a semantic point of view.

Thus, to obtain an empirical p-value for the test of dependence of each axis of the CA, a Monte-Carlo approach following these steps is proposed:

(i) Simulate B contingency tables by permuting the product labels of descriptions at the individual level and then compute the corresponding virtual contingency table
(ii) Perform a CA on each of the simulated contingency tables
(iii) Compute all $Q_k (1 \leq k \leq D)$ statistics for each of the simulated contingency tables
(iv) Compute the p-value of each Q_k as: $\frac{1 + \sum_{s=1}^{B} I(Q_{k_S} \geq Q_{k_{obs}})}{1 + B}$

where I is the identity function equal to 1 when its argument is true and 0 otherwise, B is the number of simulations (set to 1000 in following examples), $Q_{k_{obs}}$ is the observed statistic at step k, Q_{k_s} is the S-th $(1 \leq S \leq B)$ statistic at step k computed from the simulations and 1 stands for the observed contingency table (Davison & Hinkley, 1997).

The permutation procedure proposed here is the same one as the one proposed by Meyners et al. (2013) and is similar to the one proposed by Meyners and Pineau (2010) and Wakeling, Raats, and MacFie (1992).

2.2. Accounting for the dimensionality of the dependence when investigating the variability in the products' coordinates in the CA space

Performing a CA on the word-by-product contingency table does not account for the subject's variability, which means that it is impossible to assess the stability of the products' coordinates in the CA space, and thus it is impossible to know if the products are significantly discriminated. Computing the products' confidence ellipses with parametric bootstrap (Ringrose, 2012) presents two major limitations. First, it does not take into account the subjects' individual source of variation. Second, it assumes observations are independent from each other, for both products and words, which is not the case for CATA and FC data as explained in Section 2.1. This approach is thus not appropriate. The total bootstrap methodology (Cadoret & Husson, 2013) is well suited to compute confidence ellipses for the products' coordinates in a CA space. This methodology consists of generating virtual panels with random resampling with replacement of the actual panel. Then, the products' configurations of the virtual panels are rotated on the products' configuration of the actual panel thanks to Procrustes rotations. The total bootstrap methodology enables to take into account the specificity of the subjects' individual data as well as the dependence between observations. The main issue when using this methodology is to determine how many axes to take into account in the Procrustes rotations. It seems that this decision is usually arbitrary and can lead to taking into account for example two axes (Alcaire et al., 2017; Vidal et al., 2018) or four axes (Antúnez, Vidal, de Saldamando, Giménez, & Ares, 2017). The more axes one takes into account when performing the Procrustes rotations, the more degrees of freedom are available to find an optimal rotation and thus, the smaller the ellipses. Then, the decision to take into account only two axes can probably be explained by the fact that this is the most conservative option and thus protects from over-interpretation. However this practice can lead to overestimating the variability in the products' coordinates and thus to underestimating products' discrimination. It is necessary to have an objective criterion for selecting the number of dimensions of the space in which the Procrustes rotations must be performed. For that purpose, applying Procrustes rotations in the subspace generated by the significant CA axes is proposed.

2.3. Accounting for the dimensionality of the dependence when investigating relations between products and words

The two approaches presented in the introduction, the chi-square per cell and the MDA, differ in how they consider the data, but none of them considers the dimensionality of the dependence. In addition, MDA is flawed by the fact that it considers the angle between a product vector and a word vector but not their norms. Indeed, the vector norm represents the strength with which a product or a word deviates from the independence, which is crucial information that must be taken into account. To account for all the information, scalar products should be used instead of MDA. Even if the scalar products are the valid way to interpret relations between the product vectors and the word vectors in the CA space, it still has two limitations. First, the values of scalar products can be negative or positive and they are not bounded, thus they are not intuitive, difficult to interpret and can only be compared relative to each other. Second, to the best of our knowledge, there is no criterion to determine if a given scalar product is large enough to consider the association significant. Thus, the other approach, chi-square per cell, was retained. Nevertheless, this approach has some limitations. The chi-square distribution is not valid for use in this context because of the reasons evoked in introduction (Section 1) and even more because chi-square distribution is not adapted for 2×2 contingency tables (Yates, 1984). This limitation can be overcome using the Fisher's exact test (Fisher, 1935). This test has the benefit of not relying on any distribution and then requires no specific conditions to be met. The second limitation is that chi-square per cell is performed on the raw dataset and thus on all axes of dependence, which may result in accounting for axes that are just noise and thus may lead the user to over-interpret his or her data. To overcome this limitation and determine which words are the most cited for each product, the following approach is proposed:

(i) Establish the number of significant CA axes in the sense of dependence using the procedure presented in Section 2.1
(ii) Reverse the CA computations on the significant axes to compute the derived contingency table corresponding to the significant axes
(iii) Perform Fisher's exact tests per cell on the derived contingency table accounting for the significant axes

The step of reversing the CA computation on the significant axes is detailed in the Appendix.

2.4. Case study datasets

The study took place at the Centre for Taste and Feeding Behaviour, Dijon, France. Fifty-nine regular (at least once per two weeks) consumers of red wine (16 men, 43 women, 18 to 60 years old) were recruited from a population registered in the ChemoSens Platform's PanelSens database. This database has been registered with the relevant authority (Commission Nationale Informatique et Libertés—CNIL—authorisation no. 1148039). The subjects were compensated for their participation in the study. They carried out a CATA task on four French red wines from different regions: Bordeaux (Bor), Languedoc (Lan), Gamaret wine from Beaujolais (Gam) and Val de Loire (Val). For each product, the CATA task was carried out by sensory modality: visual, olfactory and gustatory. The gustatory description was itself divided into global perception and aromas. All the CATA descriptors were selected thanks to the expertise of wine professionals. The collected data were then stored in four contingency tables, one per step, by cross tabulating the citation counts of the descriptors (columns) by the products (rows).

2.5. Analyses

All analyses and computations were performed using R 3.5.1 (R Core Team, 2018). The examples are given using contingency tables collected with CATA but it is important to remember that all the presented approaches can be used with contingency tables collected with FC.

The aim of this case study is to compare the results provided by the analytical methods proposed in Sections 2.1 to 2.3 to those from methods belonging to the chi-square rationale commonly used on contingency tables. For that purpose, the results provided by the chi-square distribution and the Monte-Carlo approach for computing the p-values of the tests of dependence were compared, as well as the difference in results after performing the Procrustes rotations in the total bootstrap procedure using either the significant axes (when more than two) or the first two axes. For the tests of dependence, any p-value less than the α risk of 5% was considered significant. Ellipses of the total bootstrap procedure were computed with an α risk of 5%. In addition, the results of the use of Fisher's exact tests per cell accounting for all the axes were compared to the results from Fisher's exact tests per cell accounting for the significant axes in the sense of dependence. The Fisher's exact tests were conducted with a one-sided greater alternative hypothesis, which means that only cells with a larger observed value than the expected value were investigated. The results of these tests are presented with two different levels of α risk, namely $\alpha = 5\%$ and $\alpha = 15\%$. The motivation for this is not to miss descriptive information concerning the products. The results presented for $\alpha = 5\%$ can be considered as significant descriptions of the products while the results presented for $\alpha = 15\%$ can be considered as tendencies in the description of the products.

It is important to highlight here that the aim of the following case study was not to conduct full interpretation ending with product comparisons, but to compare only the outputs of the proposed analyses to those of the more traditional ones in order to underline the potential differences between them.

3. Results

3.1. Dependence of CA axes

Table 1 shows similar conclusions between the results provided by the chi-square distribution and the Monte-Carlo approach for the tests of dependence of the first axes. For the gustatory global perception data, regarding the other axes, the same conclusions are also provided by the two approaches: two axes are significant in the sense of dependence. In contrast, differences exist between the results provided by the chi-square distribution and the Monte-Carlo approach concerning the tests of dependence of the second and the third axes for the olfactory data and gustatory aromas data. According to the non-valid chi-square distribution, only the first axis is significant in the sense of dependence whereas the Monte-Carlo approach reveals that there are actually two significant axes in the sense of dependence for the gustatory aromas data and three significant axes in the sense of dependence for the olfactory data.

3.2. Variability in the products' coordinates in the CA space

Fig. 1 shows information in line with the tests of dependence based on the Monte-Carlo approach.

For the visual sense CA, Fig. 1 (a) shows that ellipses confirm the results provided by the Monte-Carlo approach since the ellipses' projections on the second axis strongly overlap.

For the olfactory sense CA, Fig. 1(d) shows that the third axis indeed captures some dependence and information as it isolates the product Val from the others. If the usual relative criterion of accounting for approximately 70–80% of the inertia was used for this CA, the third dimension would not have been considered and thus some information would have been lost. Further, the comparison of Fig. 1(b) and Fig. 1(c) is a great example of possible misinterpretations and missed information resulting from arbitrarily setting the number of axes to two to perform the Procrustes rotations in the total bootstrap procedure. Indeed, looking at Fig. 1(b), the product Val seems not to be different from the products Gam and Lan whereas it is indeed on the third axis as well as on the second axis when all significant axes are considered for Procrustes rotation (Fig. 1(c)). This information is taken into account when setting the relevant number of axes to perform the Procrustes rotations. Thus looking at Fig. 1(c), we can see that Val is different from Gam and Lan. This example shows the real importance of taking into account all the significant axes in the sense of dependence to perform the Procrustes rotations in the total bootstrap procedure.

For the gustatory global perception CA, ellipses also confirm the results provided by the Monte-Carlo approach since the products Val and Lan are different from the products Bor and Gam on the second axis (Fig. 1(e)).

For the gustatory aromas CA, ellipses, computed with the most conservative option, show that the second axis captures a significant dependence as two product pairs (Val vs. Bor & Val vs. Lan) are different on the second axis, while the p-values computed using the chi-square distribution suggest that this second axis is not significant. In this example, the Monte-Carlo approach, compared to the chi-square distribution, seems to be better aligned with the information provided by the ellipses.

3.3. Relations between products and words

Fig. 2 shows that using Fisher's exact tests per cell on all the axes leads to the over-interpretation of some dependent relations that are not significant. Indeed, for the visual data, when accounting for all the axes, there are tendencies for the product Gam to be more associated with the words Black and Opaque whereas when accounting only for the first significant axis, the product Gam is definitely associated with the word Violet and not associated with the words Black and Opaque. The product Val, when accounting for all the axes, is associated with the word Violet whereas when accounting only for the first significant axis, the product Val is not associated with any words. For the gustatory aromas data, when accounting for all the axes, there are tendencies for the products Gam and Lan to be more associated with the word Red fruit whereas when accounting for the first two significant axes, the product Gam is not associated with any words, and the product Lan is definitely associated with the word Red fruit. These two examples show the need to perform Fisher's exact tests per cell using only the

Table 1
P-values of the test of dependence for each axis of each correspondence analysis performed on the four contingency tables computed by either the chi-square distribution or the Monte-Carlo approach.

Sensory modality	Computation of the p-value	Chi-square/Axis 1	Axis 2	Axis 3
Visual sense	Chi-square distribution	< 0.001	0.9882	0.9403
	Monte-Carlo approach	< 0.001	0.5134	0.3016
Olfactory sense	Chi-square distribution	< 0.001	0.0545	0.5132
	Monte-Carlo approach	< 0.001	0.0019	0.0089
Global perception from the gustatory sense	Chi-square distribution	< 0.001	0.0309	0.8652
	Monte-Carlo approach	< 0.001	< 0.001	0.1448
Aromas from the gustatory sense	Chi-square distribution	0.0032	0.3378	0.8635
	Monte-Carlo approach	< 0.001	0.0069	0.2507

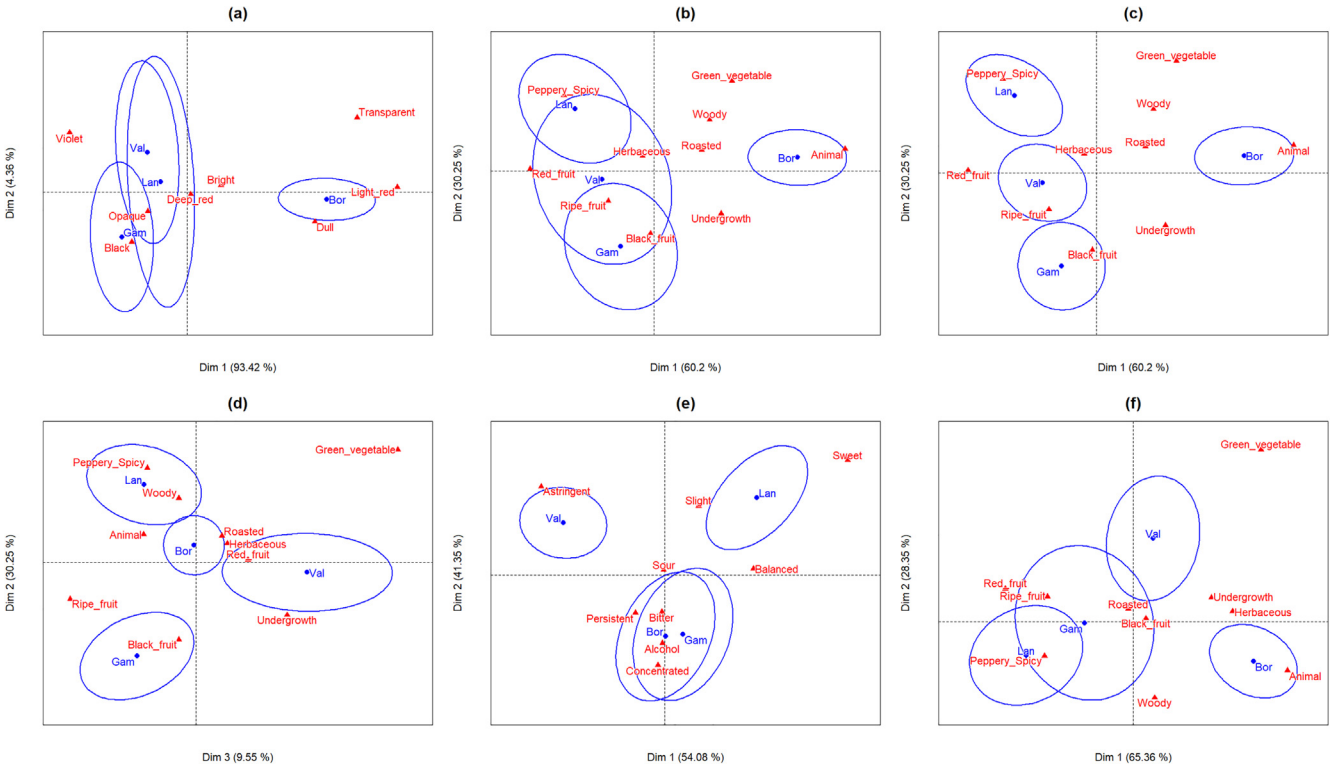


Fig. 1. Correspondence analysis of the four contingency tables with confidence ellipses computed with total bootstrap: (a) axes 1–2 of the visual sense with total bootstrap considering the first two axes, (b) axes 1–2 of the olfactory sense with total bootstrap considering the first two axes, (c) axes 1–2 of the olfactory sense with total bootstrap considering the three axes, (d) axes 3–2 of the olfactory sense with total bootstrap considering the three axes, (e) axes 1–2 of the global perception from the gustatory sense with total bootstrap considering the first two axes, (f) axes 1–2 of the aromas from the gustatory sense with total bootstrap considering the first two axes.

information provided by the significant axes in the sense of dependence. This approach prevents the user from over-interpreting some of the associations that are not sufficiently strong to be considered significant and prevents the user from missing some significant associations due to tests performed on a dataset containing noise.

The example of the global perception gustatory data shows that the differences between accounting for all the axes and accounting for the significant axes in the sense of dependence sometimes do not drastically change the conclusions. In this example, the differences are only based on some tendencies of associations.

For the olfactory data, by construction, no difference exists since all the axes are significant.

4. Discussion

The Monte-Carlo approach had a real benefit in the computation of the p-values of the chi-square test and the tests of dependence of the CA axes. Indeed, it enabled the consistent estimation of the distribution under the null hypothesis that takes into account the nature of the data. The examples presented showed that p-values computed with the Monte-Carlo approach and with the chi-square distribution do not always lead to different conclusions. However, it is common to find differences between these two approaches. As shown in the examples, the Monte-Carlo approach always provided information in line with the one provided by the confidence ellipses contrary to the chi-square distribution. This finding shows that in addition to its theoretical benefit of taking into account the nature of the data, in practice the Monte-Carlo approach also provided conclusions consistent with other information. Furthermore, in the given examples, the p-values of the Monte-Carlo approach were systematically lower than those computed using chi-square distribution. This finding suggests a higher power in dimensionality detection for the Monte-Carlo approach. Despite its benefits, the Monte-Carlo approach has a limitation: the computational time. Simulating 1000 contingency tables with the procedure explained in Section 2.1.2 takes between 10 and 20 s. If the user wants to simulate more contingency tables to better estimate the distribution under the null hypothesis, the computational time can rapidly increase.

To the best of our knowledge, testing the dependence of the CA axes has never been used in sensory and consumer research. This test is a great improvement in the analysis of contingency tables collected with CATA and FC. It enables the determination of the number of dimensions in which the dependence between products and words, if any, is large enough to be considered significant according to a statistical criterion. It prevents misinterpretations or over-interpretations and missing relevant information provided by CA axes beyond the first plan. The result of this test is also a solid basis on which further computations can rely such as the total bootstrap procedure and the investigation of associations between products and words. For the total bootstrap procedures applied on CATA and FC data, these tests are a real improvement as they provide an objective and relevant manner of determining how many axes must be taken into account for the Procrustes rotations, which prevents the user from considering two products as not being significantly discriminated when they are indeed.

Fisher's exact tests were performed with a one-sided greater alternative meaning that only observed counts that were potentially larger than the expected counts were investigated. This choice was made because of the task asked to subjects. Concerning the FC task, it is asked to subjects to describe the products in their own words. It is thus reasonable to assume that the words used to describe a product are indeed descriptive of and applicable to the product. However, assuming that because a subject does not say a given word for a product implies that this word is not applicable to the product is a very strong assumption. For the CATA task, the situation is a little different: subjects are asked to quote among a list of words, which ones apply to the products. It is thus

		Fisher's exact tests per cell on all the axes				Fisher's exact tests per cell on the significant axes			
		Bor	Gam	Lan	Val	Bor	Gam	Lan	Val
Visual sense	Violet	3	20	19	25	3	20	19	25
	Opaque	23	38	33	35	23	38	33	35
	Dull	8	3	4	3	8	3	4	3
	Light red	18	3	4	6	18	3	4	6
	Bright	33	24	30	27	33	24	30	27
	Deep red	34	34	37	35	34	34	37	35
	Black	9	19	14	15	9	19	14	15
	Transparent	9	1	4	4	9	1	4	4
		Bor	Gam	Lan	Val	Bor	Gam	Lan	Val
Olfactory sense	Black fruit	25	33	17	24	25	33	17	24
	Roasted	6	3	4	4	6	3	4	4
	Red fruit	8	19	25	26	8	19	25	26
	Green vegetable	5	0	3	4	5	0	3	4
	Peppery / Spicy	9	8	22	12	9	8	22	12
	Ripe fruit	8	13	11	7	8	13	11	7
	Animal	24	7	7	5	24	7	7	5
	Undergrowth	19	13	6	15	19	13	6	15
	Herbaceous	4	3	4	4	4	3	4	4
	Woody	21	8	16	11	21	8	16	11
		Bor	Gam	Lan	Val	Bor	Gam	Lan	Val
Global perception from the gustatory sense	Alcohol	22	26	13	17	22	26	13	17
	Slight	17	17	26	21	17	17	26	21
	Astringent	17	14	16	40	17	14	16	40
	Bitter	15	14	10	12	15	14	10	12
	Concentrated	23	21	10	14	23	21	10	14
	Balanced	24	18	26	12	24	18	26	12
	Sweet	3	7	13	3	3	7	13	3
	Persistent	29	28	17	27	29	28	17	27
	Sour	19	18	17	19	19	18	17	19
		Bor	Gam	Lan	Val	Bor	Gam	Lan	Val
Aromas from the gustatory sense	Red fruit	6	19	21	16	6	19	21	16
	Ripe fruit	6	11	12	10	6	11	12	10
	Green vegetable	5	2	1	9	5	2	1	9
	Black fruit	28	26	21	23	28	26	21	23
	Roasted	5	4	5	5	5	4	5	5
	Peppery / Spicy	13	13	25	12	13	13	25	12
	Herbaceous	7	3	3	5	7	3	3	5
	Woody	20	13	15	7	20	13	15	7
	Undergrowth	22	13	10	18	22	13	10	18
	Animal	18	6	5	7	18	6	5	7

Fig. 2. Contingency tables of the four CATA tasks. The highlighted cells show the significant results of Fisher's exact tests per cell considering all the axes and the significant results of Fisher's exact tests per cell considering the significant axes in the sense of dependence. The cells highlighted in light green are significant for $\alpha = 5\%$, and those highlighted in deep green are significant for $\alpha = 15\%$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

reasonable to assume that a descriptor that was not used to describe a product was not perceived by the subject. This can be considered a more active decision than not to cite some words in a FC task, but still the guideline was to “check-all-that-apply” and not to “not check what does not apply”. Considering these points, the decision of performing one sided greater alternative tests or two-sided alternative tests is up to the discretion of the user.

As an overall limitation, it has to be mentioned that the practical results provided through the examples arose from datasets where only four products were evaluated using CATA. The relevant results of this paper need to be confirmed on other datasets with more products and with different levels of similarity between the products.

5. Conclusion

This paper introduced a complete set of statistical tools enabling to account for the dimensionality of the dependence in contingency tables obtained with CATA and FC. First, this set includes a chi-square-based test for determining the number of significant axes in CA of a contingency table. As p-values derived from chi-square distribution are not valid in the context of contingency tables based on CATA or FC data, an alternative Monte-Carlo approach was proposed. Secondly, it was shown that the Procrustes rotations in a total bootstrap procedure to derive product confidence ellipses should be done in the subspace defined by the significant axes. Finally, to investigate which words are cited more often for each product, the paper proposed to perform Fisher’s exact tests per cell on the derived contingency table obtained by reversing the CA computation on the axes capturing significant dependence. These new tools should help the users of CATA and FC to analyse their data with more precision as the methods removed noise due to non-significant dimensions in term of dependence between products and attributes or words.

CRediT authorship contribution statement

Benjamin Mahieu: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization. **Michel Visalli:** Conceptualization, Software, Validation, Resources, Data curation, Writing - review & editing. **Pascal Schlich:** Conceptualization, Validation, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

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

Appendix:. Reversing the correspondence analysis computations

Let X be a contingency table. Performing a correspondence analysis on X consists of computing the standardised residual matrix R from X and then factorising R using Singular Value Decomposition (SVD). Factorising R using a SVD consists of writing R as follows:

$$R = UDV^{'}$$

The SVD of R is performed with weights for rows and columns equal to their respective marginal probabilities. The coordinates of the rows and the columns as well as the eigenvalues of the CA can directly be computed from U, D and V. For more details on this process and the computations, one can refer to Bock (2011).

Reversing the CA computations on the significant axes consists of computing R_{sig} as follows:

$$R_{sig} = U_{sig} D_{sig} V^{'}_{sig}$$

where U_{sig} is determined from the rows coordinates of only the significant axes, D_{sig} is determined from the eigenvalues of only the significant axes and $V^{'}_{sig}$ is determined from the columns coordinates of only the significant axes. Therefore, non-significant dependence is discarded. One critical aspect in the computations of U_{sig} and $V^{'}_{sig}$ is to determine if the software used to perform the CA returns principal coordinates or standard coordinates of the rows and the columns. U_{sig} and $V^{'}_{sig}$ have to be weighted back by the observed marginal probabilities before the computation of R_{sig}.

X_{sig} can then be computed from R_{sig} using the observed expected probabilities and the observed grand sum of X.

References

Abdi, H., & Valentin, D. (2007). Some new and easy ways to describe, compare, and evaluate products and assessors. *Proceedings of SPISE,* 5–18.

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. 7th Pangborn Sensory Science Symposium. Minneapolis, USA*.

Adery, C. A. H. (1968). A Simplified Monte Carlo Significance Test Procedure. *Journal of the Royal Statistical Society. Series B (Methodological), 30*(3), 582–598.

Agresti, A. (2007). *An Introduction to categorical data analysis* (second ed.). New York: John Wiley & Sons.

Alcaire, F., Antunez, L., Vidal, L., Zorn, S., Gimenez, A., Castura, J. C., et al. (2017). Comparison of static and dynamic sensory product characterizations based on check-all-that-apply questions with consumers. *Food Research International, 97*, 215–222.

Antúnez, L., Ares, G., Giménez, A., & Jaeger, S. R. (2016). Do individual differences in visual attention to CATA questions affect sensory product characterization? A case study with plain crackers. *Food Quality and Preference, 48*, 185–194.

Antúnez, L., Vidal, L., de Saldamando, L., Giménez, A., & Ares, G. (2017). Comparison of consumer-based methodologies for sensory characterization: Case study with four sample sets of powdered drinks. *Food Quality and Preference, 56*, 149–163.

Benzécri, J.-P. (1973). *Analyse des données. Analyse des correspondances*, *vol. 2.* Paris: Dunod.

Bock, T. (2011). Improving the display of correspondence analysis using moon plots. *International Journal of Market Research, 53*(3), 307–326.

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

Camiz, S., & Gomes, G. C. (2013). *Joint correspondence analysis versus multiple correspondence analysis: A solution to an undetected problem. Classification and Data Mining*.

Carr, B. T., Dzuroska, J., Taylor, R. O., Lanza, K., & Pansini, C. (2009). *Multidimensional Alignment (MDA): A simple numerical tool for assessing the degree of association between products and attributes on perceptual maps. 8th Pangborn sensory science symposium. Florence, Italy*.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application.* Cambridge: Cambridge University Press.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society, 98*(1).

Galmarini, M. V., Symoneaux, R., Chollet, S., & Zamora, M. C. (2013). Understanding apple consumers' expectations in terms of likes and dislikes. Use of comment analysis in a cross-cultural study. *Appetite, 62*, 27–36.

Lahne, J., Trubek, A. B., & Pelchat, M. L. (2014). Consumer sensory perception of cheese depends on context: A study using comment analysis and linear mixed models. *Food Quality and Preference, 32*, 184–197.

Lawrence, G., Symoneaux, R., Maitre, I., Brossaud, F., Maestrojuan, M., & Mehinagic, E. (2013). Using the free comments method for sensory characterisation of Cabernet Franc wines: Comparison with classical profiling in a professional context. *Food Quality and Preference, 30*(2), 145–155.

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.

Oppermann, A. K. L., de Graaf, C., Scholten, E., Stieger, M., & Piqueras-Fiszman, B. (2017). Comparison of Rate-All-That-Apply (RATA) and Descriptive sensory Analysis (DA) of model double emulsions with subtle perceptual differences. *Food Quality and Preference, 56*, 55–68.

Patefield, W. M. (1981). Algorithm AS 159: An efficient method of generating random R × C tables with given row and column totals. *Applied Statistics, 30*(1).

Perrin, L., & Pagès, J. (2009). Construction of a Product Space from the Ultra-Flash

Profiling Method: Application to 10 Red Wines from the Loire Valley. *Journal of Sensory Studies, 24*(3), 372–395.
R Core Team (2018). *R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing*.
Renter, D. G., Higgins, J. J., & Sargeant, J. M. (2000). *Performance of the exact and chi-square tests on sparse contingency tables. Conference on Applied Statistics in Agriculture*.
Ringrose, T. J. (2012). Bootstrap confidence regions for correspondence analysis. *Journal of Statistical Computation and Simulation, 82*(10), 1397–1413.
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.
Valentin, D., Chollet, S., Lelièvre, M., & Abdi, H. (2012). Quick and dirty but still pretty good: A review of new descriptive methods in food science. *International Journal of Food Science & Technology, 47*(8), 1563–1578.
Varela, P., & Ares, G. (2012). Sensory profiling, the blurred line between sensory and consumer science. A review of novel methods for product characterization. *Food Research International, 48*(2), 893–908.
Vidal, L., Ares, G., Hedderley, D. I., Meyners, M., & Jaeger, S. R. (2018). Comparison of rate-all-that-apply (RATA) and check-all-that-apply (CATA) questions across seven consumer studies. *Food Quality and Preference, 67*, 49–58.
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.
Yates, F. (1984). Test of significance for 2 × 2 contingency tables. *Journal of the Royal Statistical Society. Series A (General), 147*(3).

C. A multiple-response chi-square framework for the analysis Free-Comment data

Article published in Food Quality and Preference:

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

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*

Reference:

Mahieu, B., Schlich, P., Visalli, M., & Cardot, H. (2021). A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data. *Food Quality and Preference, 93*.

Food Quality and Preference 93 (2021) 104256

Contents lists available at ScienceDirect

Food Quality and Preference

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

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

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

ARTICLE INFO

Keywords:
Chi-square statistic
Multiple-response Correspondence Analysis (MR-CA)
Multiple-response dimensionality test of dependence
Multiple-response hypergeometric test
Analysis of multiple-response data

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 hypergeometric test 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.

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

* Corresponding authors.
E-mail addresses: benjamin.mahieu@inrae.fr (B. Mahieu), pascal.schlich@inrae.fr (P. Schlich).

https://doi.org/10.1016/j.foodqual.2021.104256
Received 22 October 2020; Received in revised form 22 March 2021; Accepted 29 March 2021
Available online 6 April 2021

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 $p \in \{1, \cdots, P\}$ 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 $d \in \{1, \cdots, D\}$ 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:

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

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

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 affects 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, $p \in \{1, \cdots, P\}$
- $\boldsymbol{c}$ a column matrix of size $D \times 1$ whose elements equals C_d/E, $d \in \{1, \cdots, D\}$
- $\boldsymbol{D_r}$ a diagonal matrix of size $P \times P$ whose diagonal elements equal E_p/E, $p \in \{1, \cdots, P\}$
- $\boldsymbol{D_c}$ a diagonal matrix of size $D \times D$ whose diagonal elements equal C_d/E, $d \in \{1, \cdots, D\}$

- $\boldsymbol{X}$ A matrix of size $P \times D$ whose general term equal $n_{pd}/E,\ p \in \{1, \cdots, P\},\ d \in \{1, \cdots, D\}$

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 $\boldsymbol{S}$, Γ the diagonal matrix of singular values of $\boldsymbol{S}$ and $\boldsymbol{V}$ the matrix of right singular vectors of $\boldsymbol{S}$ such that $\boldsymbol{S} = \boldsymbol{U}\Gamma\boldsymbol{V}^t$. Similarly to the usual CA, the principal coordinates of the products are defined as $\boldsymbol{D}_r^{-\frac{1}{2}}\boldsymbol{U}\Gamma$ 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 $p \neq p' \in \{1, \cdots, P\}$ 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 the multiple-response chi-square framework

This section transposes the methodologies from Mahieu et al. (2020a) to the multiple-response 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 $\boldsymbol{S}$, Γ_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 $\boldsymbol{S}_k = \boldsymbol{U}_k\Gamma_k{\boldsymbol{V}_k}^t$. 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:

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

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 $\chi^{2\ (*)}_{mrk}$ statistics, $k = 1, \cdots, K$, as $\chi^{2\ (*)}_{mrk} = E \times \sum_{i=k}^{K} \lambda_i^{(*)}$ for each of the simulated contingency tables
(iv) Compute the p-value of each $\chi^2_{mr}k$ as the proportion of $\chi^{2\ (*)}_{mrk}$ 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 total bootstrap test, 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 hypergeometric 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 hypergeometric test to test the following hypotheses for a given$p \in \{1, \cdots, P\}$ and a given $d \in \{1, \cdots, D\}$:

$H_0 : \pi_d^p = \pi_d$

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

The multiple-response hypergeometric test is based on a Monte-Carlo procedure. In this procedure, for each product $p \in \{1, \cdots, P\}$, 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 hypergeometric tests can be performed with a two-sided alternative hypothesis or a one-sided greater alternative hypothesis.

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.

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)	/

Finally, it is proposed to perform the multiple-response hypergeometric tests 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 $\boldsymbol{S}_{sig} = \boldsymbol{U}_{sig}\Gamma_{sig}\boldsymbol{V}_{sig}{}^t$ with $\boldsymbol{U}_{sig}$ the matrix of left singular vectors of $\boldsymbol{S}$ corresponding to the significant axes, Γ_{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 hypergeometric tests 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

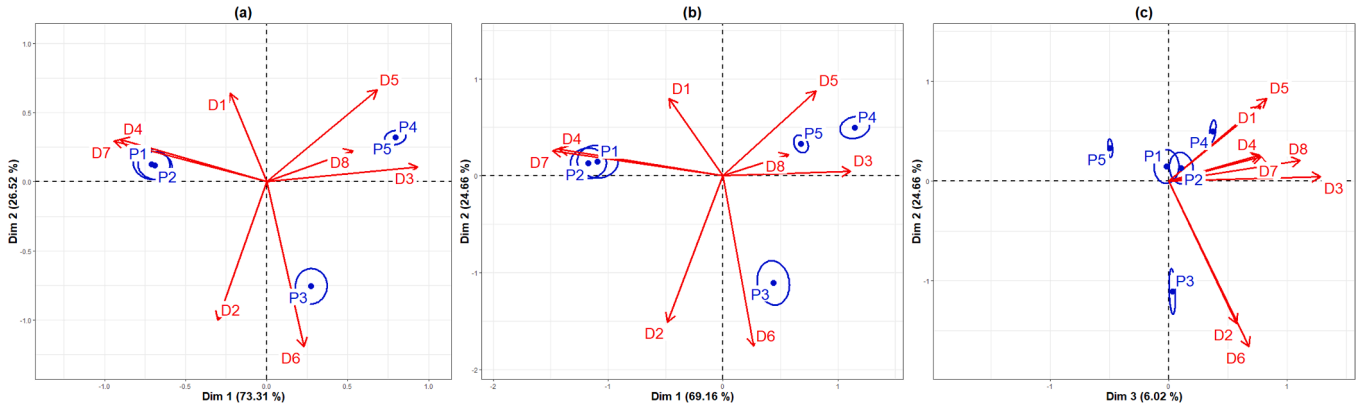


Fig 1. Biplot from Correspondence Analysis of the texture dataset: (a) usual CA (axes 1–2), (b) MR-CA (axes 1–2) and (c) MR-CA (axes 3–2).

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 hypergeometric test 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

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) & (d)). On the contrary, in usual CA, *P3* seems to be associated with *D1*, *D5,* and *D6* and slightly with *D2* (Fig. 2(a) & (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,

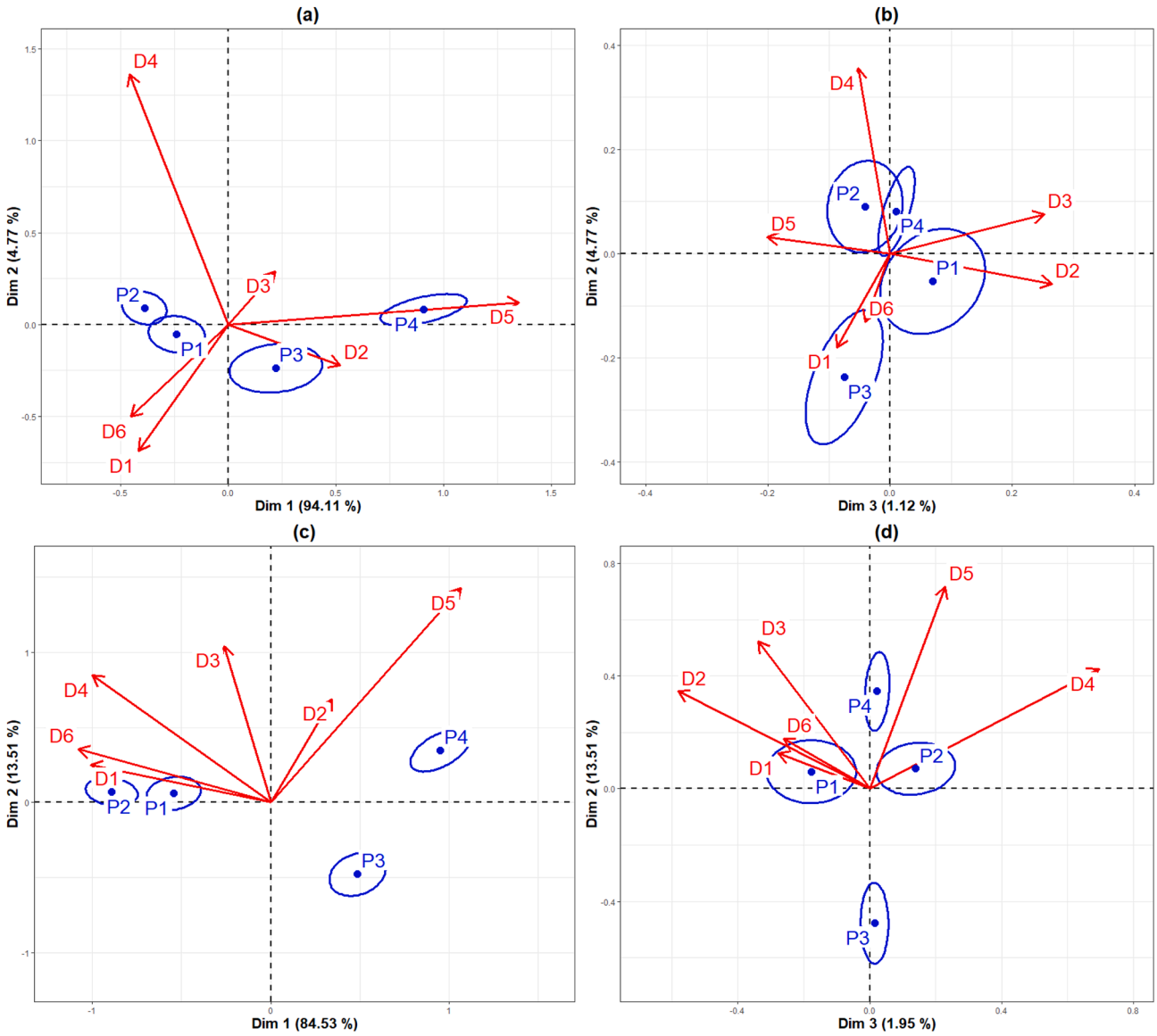


Fig 2. Biplot from Correspondence Analysis of the flavor dataset: (a) usual CA (axes 1–2), (b) usual CA (axes 3–2), (c) MR-CA (axes 1–2) and (d) MR-CA (axes 3–2).

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) & (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), (b), and (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*

		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	

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

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 hypergeometric test introduced in this paper takes 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 s 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 min 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.

CRediT authorship contribution statement

Benjamin Mahieu: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing - original draft, Writing - review & editing. **Pascal Schlich:** Conceptualization, Validation, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Michel Visalli:** Conceptualization, Software, Validation, Resources, Writing - review & editing. **Hervé Cardot:** Conceptualization, Methodology, Software, Validation, Resources, Writing - review & editing, Writing - original draft, Supervision, Project administration.

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).

Appendix:. Simulations

To investigate the benefits and/or the downsides of performing the multiple-response hypergeometric 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 $\times$ 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 $\times$ 2 levels of deviation intensity) of probabilities were thus generated. Each of the 8 matrices contains 50 cells (5 products $\times$ 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 $\times$ 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 hypergeometric 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 hypergeometric 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 2.

Table 2 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

Table 2
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.

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
0.1	*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
	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
0.2	*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
	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

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 hypergeometric 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 hypergeometric tests per cell on the derived contingency table corresponding to the significant axes.

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 ed.).

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*, 103924. https://doi.org/10.1016/j.foodqual.2020.103924

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*, 103937. https://doi.org/10.1016/j.foodqual.2020.103937

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*, 104159. https://doi.org/10.1016/j.foodqual.2020.104159

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, Ian N., Raats, Monique M., & MacFIE, Halliday 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.

Chapter IV:
Performances of Free-Comment as compared to Check-All-That-Apply

A. Context and contents

As mentioned in Chapter I, the ability of Free-Comment (FC) to differentiate and characterize a set of products no longer needs to be demonstrated. Further, its ability to provide in some applications similar product configuration and product characterizations to conventional sensory profiling suggests the descriptive sensory information provided by FC is not flawed. However, FC applications remain few and FC has never been compared to other descriptive methods of sensory analysis with consumers in terms of product discrimination and characterization. In addition, the stability of the descriptive sensory information provided by FC has never been investigated. This chapter proposes to remedy these limitations by comparing the performances of FC to those of Check-All-That-Apply (CATA) in terms of product discrimination and characterization as well as in terms of stability of the provided descriptive sensory information. The decision of challenging FC against CATA was taken because, as shown in chapter I, CATA is the method of reference for descriptive sensory analysis with consumers based on a presence/absence rationale without rating or ranking.

Section B proposes to compare FC and CATA in terms of product discrimination and characterization. For this comparison, 120 regular consumers of red wines were split into two groups of 60 consumers balanced in terms of age repartition and gender. The 120 consumers evaluated at home the same four commercialized red wines coming from different terroirs but depending on the group they belonged to, they reported their perception using either FC or CATA. For both FC and CATA, consumers carried out the evaluations by sensory modality in the following order: visual, olfactory, and gustatory. Results showed that FC provided better discrimination and richer characterization of products than CATA (Table 1, Figures 1 and 2 of section B). The overall dependence between

products and sensory descriptors was larger in FC than with CATA, providing a clearer differentiation of the products (Table 1 and Figure 2 of section B). For the olfactory modality, CATA did not discriminate the products (no significant axis) while FC did (two significant axes out of three) (Table 1 of section B). Finally, FC provided richer, more precise and more product-specific characterizations (Figure 3 of section B). This comparison provides another successful application of FC for descriptive sensory analysis with consumers. It further suggests that in some applications, FC can perform better than CATA and thus that FC is worth being given more attention than currently for descriptive sensory analysis with consumers.

Section C proposes to compare FC and CATA in terms of the stability of the descriptive sensory information they provide and it is also the first study of the stability of FC. For this comparison, the previous data on the red wines were used together with the data from another study. In this latter study, 147 regular consumers of milk chocolates constituted a FC group of 77 consumers and a CATA group of the remaining 70 consumers. The two groups were balanced in terms of age repartition and gender. The 147 consumers evaluated at home (FC) or at lab (CATA) the same four milk chocolates (commercial and prototypes) reporting their perception with the method of the group they belonged to. For both FC and CATA, consumers carried out the evaluations by sensory modality in the following order: texture in mouth and flavor in mouth. The size of product differences for a given dataset (one product type and one sensory modality) was estimated as the amount of overall dependence between the products and the sensory descriptors of the corresponding sensory modality. These overall dependences were compared from one dataset to another to obtain a relative ranking of the datasets in terms of the size of product differences (Table 2 of section C). The stability of both FC and CATA was investigated by comparing the outputs of bootstrapped virtual panels of different sizes to those of the actual

panel. Three aspects of the outputs were investigated: product configuration (Figure 1 of section C), joint product by descriptor configuration (Figure 2 of section C), and product by descriptor significant associations (Figure 3 of section C). Note that, to the best of our knowledge, investigating the stability of the product by descriptor information in the context of sensory analysis with consumers had never been addressed in the literature. Results showed FC at least as stable as CATA, if not better, for the three investigated aspects. For both FC and CATA, the overall level of stability increased with panel size and was highly dependent on the size of the differences between the products. The product configuration was more stable than the joint product by descriptor configuration, which was more stable than the product by descriptor significant associations. These results suggest that FC outputs are on the same level of stability as those of CATA and reinforces that FC worth being given more attention than currently in sensory analysis.

B. Discrimination and characterization of the products

Article published in Food Quality and Preference:

Free-comment outperformed check-all-that-apply in the sensory characterisation of wines with consumers at home

Benjamin Mahieu[a,*], Michel Visalli[a], Arnaud Thomas[b], Pascal Schlich[a]

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

[b] *SensoStat, Dijon, France*

Reference:

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

Food Quality and Preference 84 (2020) 103937

Contents lists available at ScienceDirect

Food Quality and Preference

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

Free-comment outperformed check-all-that-apply in the sensory characterisation of wines with consumers at home

Benjamin Mahieu[a,*], Michel Visalli[a], Arnaud Thomas[b], Pascal Schlich[a]

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

ARTICLE INFO

Keywords:
Open-ended questions
Textual data analysis
Sensory extrinsic information
Home Used Test (HUT)
Total bootstrap test

ABSTRACT

Check-All-That-Apply (CATA) is a popular method used for collecting word-based sensory descriptions from consumers. Free-Comment (FC), as a response to open-ended questions, is an interesting alternative because it removes biases due to the use of a predefined list of descriptors. In the context of a home used test (HUT), FC enables subjects to express themselves more naturally. The present study investigated the relevance of the use of FC at home for word-based sensory description of a set of products. Two groups of 60 consumers of red wines characterised four French red wines from different terroirs performing either a CATA task or a FC task. The two sensory tasks were performed at home according to sensory modality: visual, olfactory and gustatory. The first objective was to investigate whether a FC protocol can be successfully conducted at home and whether it enables the characterisation and discrimination of a set of products. The second objective was to investigate whether extrinsic sensory information affects FC descriptions. The third objective was to investigate whether CATA and FC provide comparable information in the HUT context. The results show that an FC protocol is feasible at home and that the extrinsic sensory information did not affect FC descriptions. FC enabled better characterisation and discrimination of the products than CATA. A new test of product differences based on the total bootstrap procedure was proposed to compare FC and CATA.

1. Introduction

Sensory descriptive analysis (DA) (Brandt, Skinner, & Coleman, 1963; Cairncross & Sjostrom, 1950; Meilgaard, Civille, & Carr, 1991; Murray, Delahunty, & Baxter, 2001; Stampanoni, 1993; Stone, Sidel, Oliver, Woolsey, & Singleton, 1974) has proved itself as a high-performing tool in characterising and quantifying the sensory properties of different products. It is still extensively used for several different goals such as product development, product comparison, quality control, understanding consumer preferences, etc. However, DA requires a trained panel, which presents some practical limitations: it is expensive and time consuming because of the training phase, and it is not necessarily representative of consumers' sensory perception (Ares & Varela, 2017; Delgado & Guinard, 2011; Ramirez, Hough, & Contarini, 2001). In recent years, new, more consumer-oriented methods have emerged to overcome DA's limitations (Valentin, Chollet, Lelièvre, & Abdi, 2012; Varela & Ares, 2012). These methods can be classified into three categories: verbal-based methods, similarity-based methods and reference-based methods (Valentin et al., 2012; Varela & Ares, 2012). Among verbal-based methods, Check-All-That-Apply (CATA) (Adams, Williams, Lancaster, and Foley (2007)) is one of the most popular.

During a CATA task, the subjects are asked to choose among a list of descriptors, those that apply to a given product. Most of the time, the entire list of descriptors is presented to the subjects, but alternatively, descriptors can be presented sequentially with forced-choice questions (Jaeger et al., 2014) or in different sub-lists (Ares et al., 2013). CATA has proven itself an efficient method for the characterisation and discrimination of a set of products with consumers (Oppermann, de Graaf, Scholten, Stieger, & Piqueras-Fiszman, 2017; Valentin et al., 2012; Varela & Ares, 2012). However, the step consisting of establishing a list of descriptors is very tedious and critical for the relevance of the collected data as it may affect the results of the study (Ares et al., 2013; Hughson & Boakes, 2002). Furthermore, several sources of bias induced by the use of a predefined list of descriptors have been reported in the literature. The list influences the subjects by suggesting descriptors that they would not think about otherwise (Coulon-Leroy, Symoneaux, Lawrence, Mehinagic, & Maitre, 2017; Kim, Hopkinson, van Hout, & Lee, 2017; Krosnick, 1999). Since the list contains only a limited number of descriptors, subjects may select descriptors that are close to what they perceive but not representing exactly what they actually

* Corresponding author.
E-mail address: benjamin.mahieu@inrae.fr (B. Mahieu).

https://doi.org/10.1016/j.foodqual.2020.103937
Received 17 February 2020; Received in revised form 20 March 2020; Accepted 22 March 2020
Available online 25 March 2020
0950-3293/

perceive (Krosnick, 1999). The first descriptors of the list (in the sense of presentation order) have a greater chance of being selected (Castura, 2009; Kim et al., 2017; Krosnick, 1999). In addition, the investment of the subjects is low when performing CATA, causing them to give quick answers and not pay attention to all descriptors (Krosnick, 1999; Sudman & Bradburn, 1982; Varela & Ares, 2012). This latter issue can be addressed using Yes/No questions (Jaeger et al., 2014). However, a protocol based on Yes/No questions presents the drawback to be more time-consuming. (Meyners & Castura, 2014; Smyth, Dillman, Christian, & Stern, 2006).

Free-Comment (FC) (ten Kleij & Musters, 2003), as a response to open-ended questions, is an alternative to CATA in collecting word-based sensory descriptions. For each evaluated product, subjects are asked to describe the product in their own words without any form of restriction (Hanaei, Cuvelier, & Sieffermann, 2015; ten Kleij & Musters, 2003). Depending on the aim of the study, descriptions can be hedonic-oriented (Lahne, Trubek, & Pelchat, 2014; Luc, Lê, & Philippe, 2020; Symoneaux, Galmarini, & Mehinagic, 2012) or not. More rarely, a limitation on the number of words they can use is imposed on subjects (Ares, Giménez, Barreiro, & Gámbaro, 2010). FC has proven itself an efficient method in characterising and discriminating sets of products both with consumers and with experts (Lahne et al., 2014; Lawrence et al., 2013; ten Kleij & Musters, 2003). As FC does not require a pre-defined list of descriptors, all the above-mentioned CATA biases are no longer an issue. Besides, the investment of subjects in the sensory task may be improved if no list of descriptors is proposed. In addition, it is hypothesised that FC could be better suited than CATA for the home used test (HUT) context since it enables subjects to express themselves more naturally.

To the best of our knowledge, results from an FC protocol conducted at home have never been reported. Furthermore, the results provided by CATA and FC have never been compared.

The present study first investigated whether an FC protocol can be successfully conducted at home and whether it enables the characterisation and discrimination of a set of products. Second, due to the HUT context, the study had to investigate whether extrinsic sensory information displayed on the label of the products could affect the FC descriptions. Third, the study investigated whether CATA and FC conducted at home provide comparable information in terms of the characterisation and discrimination of a set of products.

To compare FC to CATA, the methodology presented in (Mahieu, Visalli, & Schlich, 2020) was used. In addition, a method was proposed based on the total bootstrap procedure for testing whether two products are significantly different.

2. Material and methods

2.1. Participants

To create a situation as close as possible to an everyday consumption situation, the study took place at home with 120 naïve subjects (64 men and 56 women), 18 to 60 years old. Subjects were recruited from a population registered in the ChemoSens Platform's PanelSens database. This database has been declared to the relevant authority (Commission Nationale Informatique et Libertés—CNIL—n° d'autorisation 1148039). The subjects recruited were consumers of red wines at least once every two weeks and were allocated to two groups of 60 subjects. The two groups were balanced in terms of age repartition and gender. The first group performed a CATA task while the second group performed a FC task. The bottles of wine they had to taste were the rewards of the study.

2.2. Products

Four commercialised French red wines from different terroirs were used as products for this study. The four terroirs were Bordeaux (Bor), Gamaret wine from Beaujolais (Gam), Languedoc (Lan) and Val de Loire (Val). The wines were selected from different terroirs to ensure different sensory characteristics across the products. The wines were delivered to the subjects in their respective commercial glass bottles. For the products Gam and Val, commercial labels and back labels were removed from the bottles. For the products Bor and Lan, the bottles were delivered to the subjects with their respective commercial labels and back labels. The purpose of this was to assess whether some subjects simply copy the sensory description present on the back label of the wine. This is an important point, since HUTs could occur in the presence of product labels if a study aims to compare products in their commercial packaging.

2.3. Data acquisition

2.3.1. General procedure

The subjects participated in four home-based sessions on their own computers, tablets or smartphones running TimeSens© software 2.0 (INRA, Dijon, France). To access the sessions, subjects simply had to click on a link sent to them by e-mail. Each session corresponded to the evaluation of only one product and lasted approximately 10 min. The minimum interval between two sessions was forced to be at least 24 h, and an average of 72 h was observed. The subjects were invited to preserve and consume the wines in the manner they usually do. Depending on the group to which the subjects belonged, they were asked to perform a CATA or a FC task.

2.3.2. CATA task

For each product, the CATA task was carried out by sensory modality in the following order: visual, olfactory and gustatory. The gustatory description was itself divided into global perception and aromas. The following CATA descriptors were selected according to the expertise of wine professionals:

- Visual sense: *Violet, Opaque, Dull, Light_red, Bright, Deep_red, Black, Transparent*
- Olfactory sense: *Black_fruit, Roasted, Red_fruit, Green vegetable, Peppery_Spicy, Ripe_fruit, Animal, Undergrowth, Herbaceous, Woody*
- Global perception from the gustatory sense: *Alcohol, Slight, Astringent, Bitter, Concentrated, Balanced, Sweet, Persistent, Sour*
- Aromas from the gustatory sense: *Red_fruit, Ripe_fruit, Green_vegetable, Black_fruit, Roasted, Peppery_Spicy, Herbaceous, Woody, Undergrowth, Animal*

The descriptors were presented in a different randomised order for each subject but with a constant order across evaluations for a given subject. For each of the four above-mentioned steps, the following instruction was given to the subjects: “Check in the subsequent list the words that apply to this wine”.

2.3.3. FC task

For each product, the FC task was carried out by sensory modality in the following order: visual, olfactory and gustatory. For each of the three steps, the following instructions were given to the subjects:

- Visual sense: “Describe the visual characteristics of the wine”
- Olfactory sense: “Describe the olfactory characteristics of the wine”
- Gustatory sense: “Describe the gustatory characteristics of the wine”

No particular restriction was given to the subjects on the manner of stating their descriptions.

2.4. Data treatment

2.4.1. CATA data

To facilitate the comparison with FC, the two types of data from the

Step	Fictive description	Automated or manual	Number of different words related to: visual sense	Number of different words related to: olfactory sense	Number of different words related to: gustatory sense
0 (Original description)	The wine is dry and fruity (peaches) but not sweet. Violet aromas can be found and a little of bitterness. The taste in mouth is pleasant!	/	484	576	798
1 (Cleaning)	the wine is dry and fruity peaches but not sweet violet aromas can be found and a little of bitterness the taste in mouth is pleasant	automated	484	576	798
2 (Lemmatisation)	the wine is dry and fruity peach but not sweet violet aroma can be find and a little of bitterness the taste in mouth is pleasant	automated	405	455	634
3 (Accounting for negations)	the wine is dry and fruity peach but not_sweet violet aroma can be find and a little of bitterness the taste in mouth is pleasant	automated	429	503	729
4 (Keeping only nouns and adjectives)	wine dry fruity peach not_sweet violet aroma bitterness taste mouth pleasant	automated	138	146	225
5 (Removing hapax per product)	wine dry fruity not_sweet violet aroma bitterness taste mouth pleasant	automated	89	92	143
6 (Removing uninformative words)	dry fruity not_sweet violet bitterness	manual	30	37	45
7 (Merging nouns and adjectives)	dry fruity not_sweet violet bitter	manual	22	32	35
8 (Grouping of words)	dry fruity_blackberry not_sweet violet_floral bitter	Classification (automated) Grouping (manual)	13	23	25
9 (Removing low cited words per product)	dry fruity_blackberry violet_floral bitter	Automated	12	14	20

Fig. 1. Evolution of a fictive description throughout FC data treatment and number of words remaining after each treatment for each sensory modality in the study of the paper.

gustatory sense (global perception and aromas) were merged together and treated as gustatory data. The data from each of the three sensory modalities (visual, olfactory and gustatory) were treated separately.

The number of times each descriptor was checked for each product was computed at the panel level. Then, the corresponding contingency table containing the citation counts of each descriptor for each product was built.

2.4.2. FC data

As FC descriptions were collected in French, all subsequent treatments were performed in French. The retained words resulting from the treatments were then translated to English for the present paper.

All FC data treatments were performed using R 3.5.1 (R Core Team, 2018). The lexicon provided with IRaMuTeQ© software (Ratinaud, 2014) was used for lemmatisation and part-of-speech tagging. The data from each of the three sensory modalities were treated separately.

Each action performed throughout the FC data treatment and presented subsequently are associated with a step number in the text. Fig. 1 shows the evolution of a fictive description throughout FC data treatment, and the number of words remaining after each treatment for each sensory modality. The second column of Fig. 1 shows the evolution of the fictive description throughout each step of FC data treatment. The third column of Fig. 1 indicates whether the step is automated or manual. “Manual” refers to a manual intervention in the R code while “automated” refers to no intervention. Columns four to six of Fig. 1 gives the number of remaining words after each step of FC data treatment for each of the three sensory modalities (visual, olfactory and gustatory).

2.4.2.1. Lemmatisation and filtration. The corpus of descriptions was cleaned (step 1) and lemmatised (step 2). The negations (e.g. “not something”, “not very something”, etc.) were considered integral words (step 3). All grammatical classes other than nouns and adjectives were removed from the corpus (step 4). The number of times each word was cited for each product was computed at the panel level. Every word not cited at least two times for at least one same product was removed from the corpus (step 5). All uninformative words (e.g. “wine”, “visually”, “mouth”, etc.) and hedonic words (e.g. “pleasant”, “good”, etc.) were removed from the corpus (step 6). The noun and adjective forms of a word (e.g. “bitterness” and “bitter”) were grouped together and considered a single word (step 7).

2.4.2.2. Grouping of words. The grouping of words step corresponds to the step 8 of Fig. 1.

As the words used to describe the same sensory dimension can differ from one subject to another, there is a need for grouping words depicting the same information. To avoid any over-grouping or subjective grouping of words (Ares et al., 2010; Lahne et al., 2014), a semi-automatic methodology was performed. For words to be grouped together they had to share similar profiles and, if that was the case, the words had to have similar semantic meanings in the study context.

Thus, after having performed all the above steps (section 2.4.2.1),

an ascending hierarchical classification of the words was performed using the methodology presented in Greenacre (1988). At each step of the classification, the segments were aggregated to keep the chi-square statistic of the collapsed contingency table as high as possible. The classification tree was cut at the step where the collapsed contingency table had the most significant chi-squared test (i.e. the lowest p-value). Among each segment of words derived from the classification, words conveying the same semantic information were grouped by the experimenter into a latent word containing all its constituting words. For example, for the olfactory sense, the words *blackberry* and *black fruit* were in the same segment and were grouped into the latent word *black_fruit_blackberry*. These grouping decisions were validated by a second experimenter. No word grouping was made between two words belonging to different segments, i.e. with too different profiles. When all consistent word grouping in terms of semantic were made in each segment, another step of classification and word grouping was performed. This procedure was repeated until no consistent word grouping in terms of semantic could be performed in a segment derived from the classification. In practice, no more than two steps of classification/grouping were necessary for the present datasets.

2.4.2.3. Building of the contingency table. Among all words and latent words (simply called words hereafter for simplification) derived from the previous procedure, only those mentioned by at least 5% of the panel for at least one same product were retained for further analysis (step 9) (Bisconsin-Júnior et al., 2020; Rios-Mera et al., 2019; Symoneaux et al., 2012).

Finally, the number of times each remaining word was cited for each product was computed at the panel level. Then, the corresponding contingency table containing the citation counts of each word for each product was built.

2.5. Data analyses

All analyses were performed using R 3.5.1 (R Core Team, 2018).

2.5.1. Evaluation of the impact of back labels on FC descriptions

This evaluation was only performed for FC descriptions and not for CATA because CATA descriptors were not used in the back labels.

For the two products with bottle back labels, the numbers of words from the back labels found in the individual FC descriptions were computed. For these computations only, FC descriptions were only cleaned and lemmatised (steps 1 and 2 of Fig. 1). Back labels were also cleaned and lemmatised before doing the computations.

2.5.2. Contingency tables analyses

A chi-square test following the Monte Carlo approach (simulations = 1000, α = 5%) presented in Mahieu et al. (2020) was performed to investigate the significance of the dependence between products and words. If the chi-square test was significant, a correspondence analysis (CA) was applied to the contingency table. The standard CA biplot (Greenacre, 2006) was used to display the CA results. This representation has the benefits that the lengths of the vectors of words are positively related to their contribution to the CA axes inertias and thus account for the ponderation used by the CA. The number of significant CA axes was determined using the Monte-Carlo tests of dependence (simulations = 1000, α = 5%) presented in Mahieu et al. (2020). The phi-square index was computed on the significant axes. The phi-square index measures the dependence between rows and columns of a contingency table. It is bounded between zero (independence) and $(k - 1)$, where k is the number of rows or the number of columns, whichever is smaller (full dependence, corresponding to a diagonal contingency table). It was used as an overall measure of associations between products and words. The confidence ellipses for the products' coordinates in the CA space were computed with a total bootstrap procedure (bootstrap samples = 1000, α = 5%) in which Procrustes rotations were performed on the significant axes (Cadoret & Husson, 2013; Mahieu et al., 2020). To assess product discrimination, a total bootstrap test (explained in the next section) was performed for each pair of products on the significant axes. To assess relations between products and words, Fisher's exact tests per cell with one-sided greater alternative hypothesis were conducted on the derived contingency table corresponding to significant axes (Mahieu et al., 2020).

2.5.3. The total bootstrap test

This test investigates the significance of product discrimination for each pair of products by using bootstrap samples. For this purpose, a canonical discriminant analysis (same as Linear Discriminant Analysis and a particular case of Canonical Variate Analysis following a One-Way MANOVA model) considering only the significant CA axes is performed for each pair of products to find the direction of maximal discrimination between the two tested products. This direction constitutes an axis, which is a linear combination of the significant CA axes. All bootstrap replicates of the two tested products are projected onto this axis of maximal discrimination. For each bootstrap sample, the difference of the coordinates on this axis of the two tested products is computed. Then, the 100-α% (α = 5% for this paper) confidence interval of the distribution of these differences (one per bootstrap sample) is computed. If the confidence interval does not contain zero, the two products are considered significantly different.

3. Results

3.1. Evaluation of the impact of back labels on FC descriptions

The products Bor and Lan had six and seven descriptive words on their back labels, respectively. Among the 120 descriptions of these products, 72% had no words in common with the back labels, 21% had one word in common and 7% had two words in common. Therefore, the presence of back labels does not seem to have affected the descriptions of the products.

3.2. Contingency tables of CATA and FC

Table 1 shows that both FC and CATA presented three axes of significant dependence for the visual and gustatory senses. For the olfactory sense, FC presented two significant axes, while CATA presented none. In addition, Table 1 shows that the intensity of the dependence (phi-square index) between products and words was systematically higher with FC than with CATA.

According to the total bootstrap test, FC and CATA enabled all product pairs to be significantly discriminated in terms of visual sense and gustatory sense. For the olfactory sense, FC enabled four product pairs out of six to be significantly discriminated, while CATA enabled no discrimination, since the dependence on olfactory information was not significant.

Fig. 2(a) and Fig. 2(b) shows that for the visual sense, all the products were discriminated on the two first axes for both CATA and FC.

Table 1
. P-values of the test of dependence for each axis and the phi-square index of each contingency table.

Sensory modality	Sensory method	P-value: chi-square / axis 1	P-value: axis 2	P-value: axis 3	Phi-square
Visual sense	*CATA*	< 0.001	0.0349	0.0029	0.0989
	FC	< 0.001	< 0.001	< 0.001	0.1697
Olfactory sense	*CATA*	0.2207	/	/	/
	FC	0.0029	0.0059	0.0729	0.1577
Gustatory sense	*CATA*	< 0.001	< 0.001	< 0.001	0.0584
	FC	< 0.001	< 0.001	< 0.001	0.1997

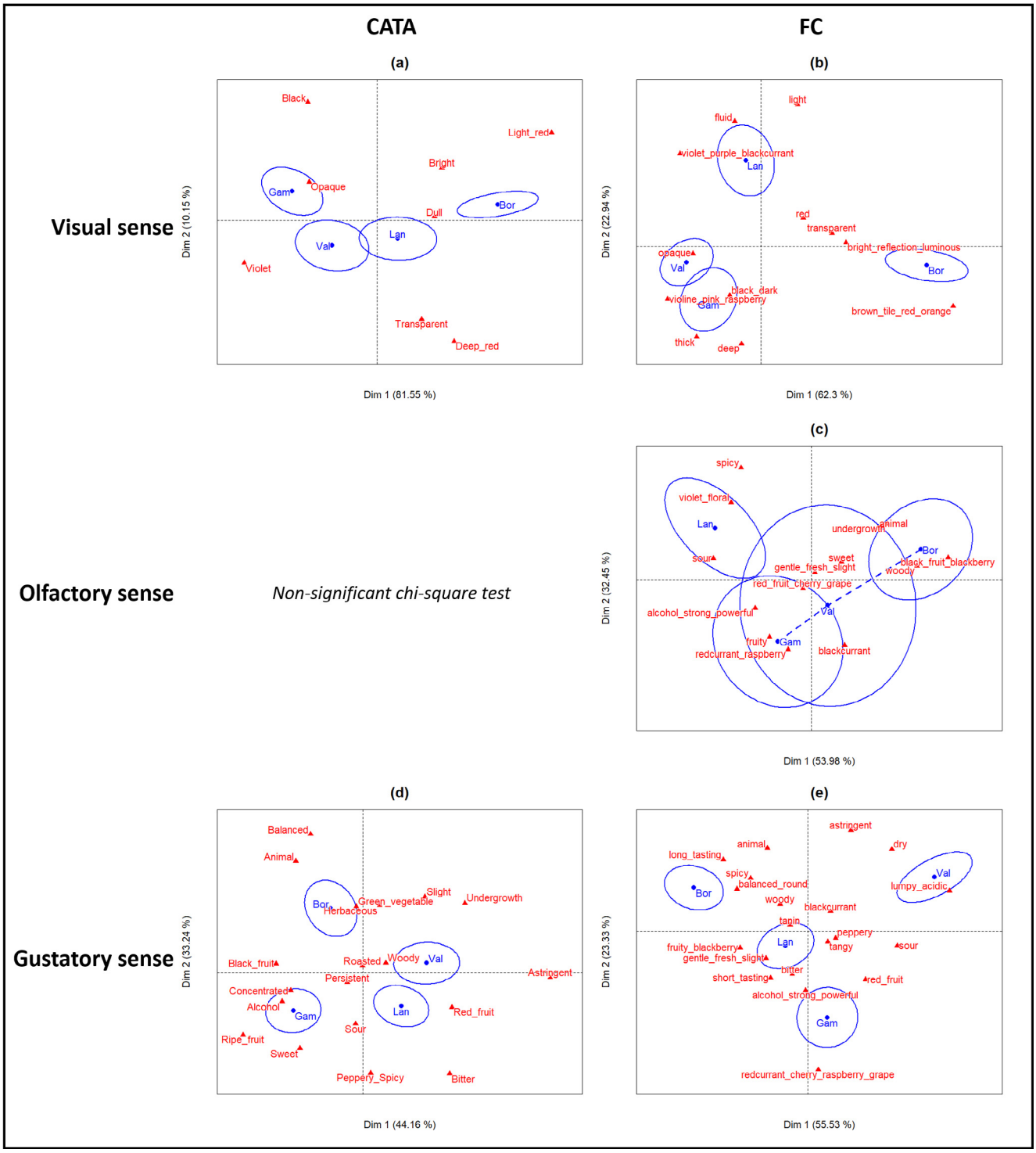


Fig. 2. Standard biplots of correspondence analyses first plans: (a) CATA, visual sense, (b) FC, visual sense, (c) FC, olfactory sense, (d) CATA, gustatory sense and (e) FC, gustatory sense. Two products linked by a dashed line are not significantly different (total bootstrap test, $\alpha = 5\%$).

Both in CATA and FC the first axis opposed Bor to Val and Gam with Lan being comprised between Bor and Val and Gam. In CATA Fig. 2(a) shows that the second axis slightly opposed Val and Lan to Gam and Bor. In FC, Fig. 2(b) shows that the second axis highly opposed Lan to the other products. In CATA, the third axis opposed Lan to Val. In FC the third axis opposed Gam to Val. In CATA, the products were positioned in pairs along the axes while in FC each axis opposed one product to the others. Both in CATA and FC, the first axis showed a gradient of colour hue associated with a gradient of transparency. In FC only, a gradient of brightness was also present on this first axis. In CATA, the second axis opposed extreme colours to middle colours of the gradient and presented a gradient of transparency. In FC, the second axis showed a gradient of colour lightness as well as a gradient of density. The third axis did not show an obvious interpretation in both CATA and FC.

For the olfactory sense, Fig. 2(c) shows that FC enabled the following pairs of products to be significantly discriminated: Lan vs Gam, Lan vs Val, Lan vs Bor and Gam vs Bor. Fig. 2(c) shows that the first axis

opposed Bor to Lan while the second axis opposed Gam to Lan and Bor. The position of Val was very uncertain. The first axis showed an opposition between different aromas as well as a gradient of powerfulness. The second axis showed a gradient of fruity.

For the gustatory sense, Fig. 2(d) and Fig. 2(e) shows that all the products were discriminated on the first two axes for both CATA and FC. In CATA, Fig. 2(d) shows that the first axis opposed Val and Lan to Bor and Gam while the second axis opposed Bor to the other products. The third axis opposed Lan and Bor to Gam and Val. In CATA, Fig. 2(e) shows that the first axis opposed Val and Bor while the second axis opposed Gam to the other products. The third axis opposed Lan the other products. In CATA, the products were positioned in pairs along the axes while in FC each axis opposed one product to the others. In CATA, all the axes opposed several combinations of aromas and global perception sensations. In FC, the first axis opposed texture sensation to aromas. The second axis showed a gradient of fruity as well as a gradient of powerfulness. The third axis did not show an obvious interpretation.

Based on visual examination of Fig. 2, the products were better separated with FC compared to CATA. This is in line with the results of the dependence intensities (phi-square index) provided by Table 1.

For visual sense, Fig. 3 shows that CATA and FC provided some similar and some different associations. The associations that differed between CATA and FC for this sensory modality can be broken into four subcategories: contradictions, words in FC that are more specific than the CATA descriptors, differences due to additional information provided by CATA and differences due to additional information provided by FC. There was one similar association between Bor and *Bright* in CATA and between Bor and *bright_reflection_luminous* in FC. There was also a similar association between Gam and *Opaque* and *Black* in CATA and between Gam and *deep*, *black_dark* and *opaque* in FC. Regarding contradictions, there were associations between Lan and *Deep_red* in CATA and between Lan and *light* in FC. Some different associations were due to the higher specificity of the words in FC compared to the CATA descriptors: there was an association between Bor and *Light_red* in CATA and between Bor and *brown_tile_red_orange* in FC. There was also an association between Lan and *Deep_red* in CATA and between Lan and *violet_purple_blackcurrant* in FC. In the same category, there was an association between Val and *Violet* in CATA and between Val and *violine_pink_raspberry* in FC. As additional information provided by CATA, there was an association between Gam and *Violet*. Finally, as additional information provided by FC compared to CATA, there was an association between Bor and *transparent* and *red*, an association between Gam and *thick* and an association between Lan and *fluid*. From an overall point of view concerning visual sense, it seems that a gradient exists from the colour of the product Bor to the colour of the product Gam that passes through the colours of the products Val and Lan. This gradient information was provided by both FC and CATA. However, FC provided more specific descriptions as well as additional information.

For olfactory sense, Fig. 3 shows that except for the product Val, FC provided meaningful product descriptions, while the dependence between words and products was not significant in CATA.

For gustatory sense, Fig. 3 shows that CATA and FC provided both similar and different associations. Among similar associations, Bor was associated with *Balanced* and *Animal* with CATA while it was associated with *balanced_round* and *animal* with FC. Lan was associated with *Peppery_Spicy* with CATA and with *peppery* with FC. Finally, the association between Val and *Astringent* was found by both methods. The different associations can be distinguished into two subcategories: additional information provided by CATA and additional information provided by FC. Among additional information provided by CATA, there was an association between Gam and *Sweet*, *Ripe_fruit* and *Black_fruit,* between Lan and *Astringent* and *Bitter* and between Val and *Red_fruit* and *Undergrowth*. Among additional information provided by FC, there was an association between Bor and *fruity_blackberry*, *long_tasting* and *spicy,* between Gam and *redcurrant_cherry_raspberry_grape* and *red_fruit,* between Lan and *alcohol_strong_powerful* and *short_tasting* and between Val and *lumpy_acidic*, *sour* and *dry*. From an overall point of view concerning gustatory sense, FC provided more information than CATA and a more specific one. However, some information provided by CATA was not provided by FC, thus, on the gustatory sense, concluding that FC description was richer than the CATA ones would be risky.

4. Discussion

Classical sensory tests performed in a lab are far from natural consumption conditions. Indeed, several samples from the same product category are usually consumed within a very short time period in a non-natural situation of consumption. Performing HUT and forcing the time between evaluations to be at least 24 h overcomes these issues and thus leads to a better approximation of natural consumption conditions. The observed average time of 72 h between two evaluations in this study suggests that consumers naturally let more than 24 hours pass between two different product consumptions. Of course, this result probably depends on the type of product being evaluated.

One potential drawback of HUT could be the effect of information if a study aims to compare products in their commercial packaging. Extrinsic information has been shown to affect liking (Mueller & Szolnoki, 2010; Ng, Chaya, & Hort, 2013), but not sensory characterisation (Tijssen, Zandstra, den Boer, & Jager, 2019). However, this latter conclusion was drawn from notations on continuous scales of attributes belonging to a predefined list. Delivering two bottles of wine with their respective labels enabled to assess the propensity of the subjects not to transcribe what they could read on the back labels of the wine. The results of this study support the idea that extrinsic information does not affect sensory characterisation in an open-ended question context. Indeed, the descriptions provided by the subjects were almost devoid of the words present on the wine back labels, since 72% of descriptions did not contain any words in the back-label descriptions, and a maximum of two words in common out of six was found for only 7% of the descriptions. However, the impact of extrinsic information in an open-ended question context may depend on the nature of the extrinsic information; thus, further investigation should be conducted to confirm the results of this study.

To the best of our knowledge, this is the first time in the context of a FC protocol that the grouping of words depicting the same information has been made based on the chi-square distance. This important step in FC data analysis is usually performed only on a semantic basis (Ares et al., 2010; Lahne et al., 2014; ten Kleij & Musters, 2003). Creating latent words enables a better understanding of the sensory characterisation provided by the subjects. This avoids discarding shades of different words conveying the same information. Furthermore, it adds transparency to the performed grouping of words, as all words that are grouped together are explicitly displayed in their corresponding latent words. The automation of the FC data treatment has the real benefit of saving time. Indeed, all the steps of the FC data treatment other than steps 6 (removing uninformative words), 7 (merging nouns and adjectives) and 8 (grouping of words) (Fig. 1) were automated. As a result, the entire process of treating FC data for the three sensory modalities, including the word-grouping step, lasted approximately one hour (given that the R code was already scripted). This is probably much faster than performing these pre-processing steps manually (Ares et al., 2010; Hanaei et al., 2015; ten Kleij & Musters, 2003). One hour may seem long compared to the almost immediate CATA data treatment but establishing the CATA list of descriptors probably takes more than one hour and may be more expensive because of pre-tests. Thus, from an overall point of view, the FC protocol is probably less time-consuming than CATA. Nevertheless, it must be mentioned that the time dedicated to the pre-treatment of FC data probably depends on the "parameters" of the study (e.g., the number of products). The time dedicated to FC data pre-treatments can be reduced throughout studies and become almost instantaneous. Indeed, lexicons can be created and enriched,

CATA

		Bor	Gam	Lan	Val
Visual sense	Violet	16.70%	50%	30%	51.70%
	Opaque	36.70%	63.30%	48.30%	53.30%
	Dull	6.70%	3.30%	8.30%	1.70%
	Light_red	16.70%	0%	5%	1.70%
	Bright	48.30%	35%	35%	36.70%
	Deep_red	61.70%	41.70%	70%	56.70%
	Black	16.70%	38.30%	25%	21.70%
	Transparent	6.70%	1.70%	6.70%	6.70%
Olfactory sense	*Non-significant chi-square test*				
Gustatory sense	Alcohol	43.30%	51.70%	45%	31.70%
	Slight	28.30%	15%	26.70%	25%
	Astringent	31.70%	21.70%	53.30%	48.30%
	Bitter	8.30%	16.70%	25%	20%
	Concentrated	41.70%	48.30%	41.70%	31.70%
	Balanced	43.30%	30%	20%	31.70%
	Sweet	6.70%	15%	10%	6.70%
	Persistent	43.30%	56.70%	40%	55%
	Sour	16.70%	23.30%	23.30%	18.30%
	Red_fruit	18.30%	23.30%	28.30%	33.30%
	Ripe_fruit	15%	26.70%	15%	10%
	Green_vegetable	6.70%	1.70%	5%	1.70%
	Black_fruit	51.70%	63.30%	40%	48.30%
	Roasted	21.70%	20%	23.30%	18.30%
	Peppery_Spicy	25%	36.70%	45%	26.70%
	Herbaceous	6.70%	1.70%	5%	3.30%
	Woody	15%	13.30%	16.70%	15%
	Undergrowth	23.30%	18.30%	18.30%	38.30%
	Animal	20%	10%	10%	6.70%

FC

	Bor	Gam	Lan	Val
red	66.70%	60%	70%	63.30%
bright_reflection_luminous	21.70%	13.30%	15%	8.30%
brown_tile_red_orange	13.30%	0%	0%	0%
thick	11.70%	30%	15%	26.70%
deep	33.30%	50%	30%	43.30%
black_dark	8.30%	20%	11.70%	11.70%
fluid	3.30%	3.30%	13.30%	8.30%
violet_purple_blackcurrant	10%	28.30%	38.30%	30%
transparent	21.70%	13.30%	16.70%	11.70%
light	6.70%	0%	11.70%	5%
violine_pink_raspberry	0%	3.30%	1.70%	15%
opaque	0%	10%	6.70%	5%
red_fruit_cherry_grape	38.30%	38.30%	38.30%	35%
redcurrant_raspberry	1.70%	6.70%	3.30%	8.30%
black_fruit_blackberry	16.70%	5%	1.70%	5%
gentle_fresh_slight	26.70%	23.30%	25%	21.70%
alcohol_strong_powerful	15%	23.30%	25%	25%
fruity	20%	33.30%	23.30%	18.30%
sweet	5%	1.70%	3.30%	5%
violet_floral	3.30%	3.30%	10%	0%
animal	5%	0%	1.70%	1.70%
woody	15%	5%	3.30%	8.30%
blackcurrant	10%	11.70%	5%	15%
spicy	5%	0%	15%	6.70%
undergrowth	8.30%	0%	3.30%	5%
sour	0%	3.30%	6.70%	1.70%
woody	16.70%	10%	10%	10%
red_fruit	8.30%	20%	11.70%	18.30%
fruity_blackberry	38.30%	28.30%	20%	15%
redcurrant_cherry_raspberry_grape	3.30%	15%	6.70%	3.30%
short_tasting	6.70%	6.70%	10%	1.70%
gentle_fresh_slight	38.30%	33.30%	28.30%	21.70%
alcohol_strong_powerful	23.30%	30%	35%	21.70%
lumpy_acidic	1.70%	11.70%	3.30%	25%
tangy	3.30%	5%	3.30%	5%
sour	11.70%	25%	11.70%	30%
animal	5%	0%	0%	1.70%
balanced_round	18.30%	6.70%	8.30%	5%
astringent	15%	6.70%	18.30%	25%
long_tasting	20%	3.30%	13.30%	5%
blackcurrant	3.30%	3.30%	1.70%	5%
bitter	8.30%	10%	8.30%	5%
peppery	0%	1.70%	11.70%	5%
tanin	21.70%	16.70%	23.30%	16.70%
spicy	11.70%	3.30%	3.30%	3.30%
dry	0%	0%	3.30%	8.30%

Fig. 3. Words by product percentages of citation across the panel for each sensory modality for CATA and FC. Cells highlighted show the results of Fisher's exact tests on significant axes, light green cells are significant for $\alpha = 5\%$ and deep green cells are significant for $\alpha = 15\%$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

which would make step 6 (removing uninformative words), 7 (merging nouns and adjectives) and 8 (grouping of words) (Fig. 1) faster throughout studies until full automation for a given product space. Several lexicons associated with different product spaces could be developed to make FC data pre-treatments automated and adapted to the type of product investigated. Further, if a fast pre-treatment is required and a lexicon does not exist yet for the type of product investigated, then manual steps (Fig. 1) can be avoided. Fixing an arbitrary threshold of cleaning (step 5, Fig. 1) before the classification step (step 8, Fig. 1) is a drawback of the automation process presented in this study. Indeed, some specific information provided by low-cited words is thereby discarded. For this reason, to avoid losing too much information, the threshold of cleaning before classification was set to a minimum of two citations for at least one same product in this study. Another drawback of automation is the cleaning of words that were misspelled. Fortunately, the same misspelling for a given word often recurs, and thus, some of the words were corrected because misspelled words were present after the cleaning step (step 5, Fig. 1).

The use of the total bootstrap test enables better use of the information provided by the total bootstrap procedure. Indeed, the test is based on the subspace generated by all significant axes rather than by only the first two axes. Considering the entire significant subspace is crucial because two products almost perfectly represented on the third axis will probably not appear discriminated on the first plan if the only criterion considered is the overlapping of their ellipses on the first plan. The probability of drawing faulty conclusions probably increases with the number of significant axes when only the information of the first plan is considered. The use of the total bootstrap test also brings another benefit: it avoids being too conservative when assessing product discrimination. Indeed, it is a statistical misconception to suppose that two quantities whose respective 95% confidence intervals fail to overlap are significantly different with an α risk of 5% (Goldstein & Michael, 1995). Only the confidence interval of the difference between the two quantities is of interest in assessing their statistical difference. Precisely, two products are different if zero is not included in the confidence interval of their difference. The proposed test follows the same rationale on the canonical discriminant axis. Indeed, if the confidence interval of the difference of the tested product coordinates on that axis contains zero, every other axis would provide confidence interval also containing zero and thus the two tested products are not significantly different in the significant space. If the confidence interval on the canonical axis does not contain zero, then at least one axis enables them to be discriminated, and thus, the products are different.

Overall, FC outperformed CATA in this study the p-values for the dimensionality tests were lower with FC compared to CATA for the visual sense and olfactory sense. Furthermore, the intensity of the dependence between products and words was larger in FC than in CATA, resulting in better product separation. For the olfactory sense, CATA did not enable characterization of the products or further discrimination of them, whereas FC did. The better performance of FC could be explained by two main elements. The first element is artifactual and echoes the biases mentioned in the introduction (section 1). The predefined list of descriptors does not seem exhaustive as some words seem very important for the characterisation of the products and were not present in the predefined CATA list of descriptors (e.g. *lumpy_acidic* for the gustatory sense). This is likely to occur whenever a predefined list is used. Furthermore, the results from the visual sense confirm the bias that subjects are led to select descriptors in the list even if these descriptors do not exactly reflect their perception. Indeed, the colour description of the products is more precise with FC and differs from the description provided by CATA (e.g. *Light_red* in CATA and *brown_tile_red_orange* in FC for the product Bor). This is summarised by the fact that words that are more specific are used with FC compared to CATA. The second element concerns the subjects' attitudes towards the task. The words used in FC to describe the products often have a specific profile compared to the descriptors used in CATA. With FC, it is quite common to observe words used for a one or for a few of the products only, whereas in CATA this almost never happens. More generally, words by product percentages of citation across the panel are much larger in CATA compared to FC as shown in Fig. 3. This might be a reason why FC outperforms CATA in terms of product discrimination. This attitude of checking many attributes for most products in CATA, called the acquiescence bias by some authors (Callegaro, Murakami, Tepman, & Henderson, 2015; Kim et al., 2017), tracks the tendency of subjects to agree with the response options in closed-ended questions regardless of their meaning and contents. For this reason, it is hard to know if the little additional information provided by CATA over FC is true information or if this is only related to the over acquiescence of subjects to CATA descriptors. The low phi-square indexes between the products and the words in CATA (Table 1) reinforce this line of reasoning.

The number of effective words (after pre-treatments) cited by the subjects for each sensory modality and each product ranged between zero and seven with an average of two. Based on the characterisations of the products, it seems that subjects are able to provide consistent FC data. Indeed, no contradictions or aberrations were observed among the significant associations between a product and the words describing it, i.e. no product was significantly characterised by words with opposite meanings. Furthermore, the olfactory and gustatory characterisations of the products are consistent with each other, i.e. some sensory dimensions associated with one product for the olfactory sense are also associated with the same product for the gustatory sense (e.g. *black_fruit_blackberry* and *fruity_blackberry* for the product Bor).

FC provided a rich and detailed description of the products in this study. One could argue that it is possible that some information associated with low citation proportions was due to chance and became statistically significant because these low citation proportions were tested against no citation at all. Thus, the relevance of this information would be then questionable. However it has to be quoted that a characteristic rarely cited, but systematically for a given product and not for the others suggests that this characteristic was indeed a marker of this product, although it was not obvious to be detected. On the contrary, a characteristic elicited a large number of times for every products, but statistically more for one product, is more likely to be due to chance. Further, in this study, most of the significant cells in the contingency tables were associated with proportions of citation that exceed 10% of the panel, arguing that these cells are unlikely to be irrelevant. Indeed, all sensory modalities confounded, only 5 out of 34 significant cells were associated with citation proportion below 10% of the panel. Of course, if the same study were performed again with another panel, it is likely that some words of this study would not be mentioned and some other would be mentioned instead. However, the same sensory information could still be represented under other words, leading to the same product discrimination. The stability of FC data remains an open question.

5. Conclusion

Collecting FC descriptions at home was feasible and enabled to characterise and discriminate products with consistent descriptions on the three sensory modalities: appearance, olfaction and gustation. The extrinsic sensory information supplied on the back labels of the bottles of wine does not seem to have affected the descriptions of the products provided by the subjects. All the steps of FC data pre-processing except removing uninformative words, merging nouns and adjective and grouping of words were fully automatized. For the grouping of words, a new semi-automatic methodology based on chi-square distance and the creation of latent words was presented. Product discrimination provided by FC and CATA was investigated using a new methodology: the total bootstrap test. This test enables not to restrict the information provided by the total bootstrap procedure to the first plan and to consider the entire significant space. Finally, by providing a more specific and richer characterisation as well as better discrimination of

the products, FC outperformed CATA in this study.

CRediT authorship contribution statement

Benjamin Mahieu: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization. **Michel Visalli:** Conceptualization, Software, Validation, Resources, Data curation, Writing - review & editing, Visualization. **Arnaud Thomas:** Conceptualization, Validation, Writing - review & editing. **Pascal Schlich:** Conceptualization, Validation, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

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

The authors would like to thank Robert et Marcel® and Sicarex® for providing their products.

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.

Ares, G., Giménez, A., Barreiro, C., & Gámbaro, A. (2010). Use of an open-ended question to identify drivers of liking of milk desserts. Comparison with preference mapping techniques. *Food Quality and Preference, 21*(3), 286–294.

Ares, G., Jaeger, S. R., Bava, C. M., Chheang, S. L., Jin, D., Gimenez, A., et al. (2013). CATA questions for sensory product characterization: Raising awareness of biases. *Food Quality and Preference, 30*(2), 114–127.

Ares, G., & Varela, P. (2017). Trained vs. consumer panels for analytical testing: Fueling a long lasting debate in the field. *Food Quality and Preference, 61*, 79–86.

Bisconsin-Júnior, A., Rodrigues, H., Behrens, J. H., Lima, V. S., da Silva, M. A. A. P., de Oliveira, M. S. R., et al. (2020). Examining the role of regional culture and geographical distances on the representation of unfamiliar foods in a continental-size country. *Food Quality and, Preference*, 79.

Brandt, M. A., Skinner, E. Z., & Coleman, J. A. (1963). Texture Profile Method. *Journal of Food Science, 28*(4), 404–409.

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

Cairncross, S. E., & Sjostrom, L. B. (1950). Flavor profiles: A new approach to flavor problems. *Food Technology, 4*, 308–311.

Callegaro, M., Murakami, M. H., Tepman, Z., & Henderson, V. (2015). Yes-no answers versus check-all in self-administered modes. *International Journal of Market Research, 57*, 203–223.

Castura, J. C. (2009). Do panellists donkey vote in sensory choose-all-that-apply questions? *8th Pangborn Sensory Science Symposium, July* (pp. 26–30). Italy: Florence.

Coulon-Leroy, C., Symoneaux, R., Lawrence, G., Mehinagic, E., & Maitre, I. (2017). Mixed Profiling: A new tool of sensory analysis in a professional context. Application to wines. *Food Quality and Preference, 57*, 8–16.

Delgado, C., & Guinard, J.-X. (2011). How do consumer hedonic ratings for extra virgin olive oil relate to quality ratings by experts and descriptive analysis ratings? *Food Quality and Preference, 22*(2), 213–225.

Goldstein, H., & Michael, J. R. H. (1995). The Graphical Presentation of a Collection of Means. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 158*(1), 175–177.

Greenacre, M. J. (1988). Clustering the rows and columns of a contingency table. *Journal of Classification, 5*(1), 39–51.

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

Hanaei, F., Cuvelier, G., & Sieffermann, J. M. (2015). Consumer texture descriptions of a set of processed cheese. *Food Quality and Preference, 40*, 316–325.

Hughson, A. L., & Boakes, R. A. (2002). The knowing nose: The role of knowledge in wine expertise. *Food Quality and Preference, 13*(7–8), 463–472.

Jaeger, S. R., Cadena, R. S., Torres-Moreno, M., Antúnez, L., Vidal, L., Giménez, A., et al. (2014). Comparison of check-all-that-apply and forced-choice Yes/No question formats for sensory characterisation. *Food Quality and Preference, 35*, 32–40.

Kim, I.-A., Hopkinson, A., van Hout, D., & Lee, H.-S. (2017). A novel two-step rating-based 'double-faced applicability' test. Part 1: Its performance in sample discrimination in comparison to simple one-step applicability rating. *Food Quality and Preference, 56*, 189–200.

Krosnick, J. A. (1999). Survey research. *Annu Rev Psychol, 50*, 537–567.

Lahne, J., Trubek, A. B., & Pelchat, M. L. (2014). Consumer sensory perception of cheese depends on context: A study using comment analysis and linear mixed models. *Food Quality and Preference, 32*, 184–197.

Lawrence, G., Symoneaux, R., Maitre, I., Brossaud, F., Maestrojuan, M., & Mehinagic, E. (2013). Using the free comments method for sensory characterisation of Cabernet Franc wines: Comparison with classical profiling in a professional context. *Food Quality and Preference, 30*(2), 145–155.

Luc, A., Lê, S., & Philippe, M. (2020). Nudging consumers for relevant data using Free JAR profiling: An application to product development. *Food Quality and, Preference*, 79.

Mahieu, B., Visalli, M., & Schlich, P. (2020). 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.

Meilgaard, M., Civille, G. V., & Carr, B. T. (1991). *Sensory Evaluation Techniques* (2nd edition). Boca Raton, FL: CRC Press.

Meyners, M., & Castura, J. C. (2014). Check-all-that-apply questions. In P. Varela, & G. Ares (Eds.). *Novel techniques in sensory characterization and consumer profiling.* Boca Raton, FL: CRC Press.

Mueller, S., & Szolnoki, G. (2010). The relative influence of packaging, labelling, branding and sensory attributes on liking and purchase intent: Consumers differ in their responsiveness. *Food Quality and Preference, 21*(7), 774–783.

Murray, J. M., Delahunty, C. M., & Baxter, I. A. (2001). Descriptive sensory analysis: Past, present and future. *Food Research International, 34*(6), 461–471.

Ng, M., Chaya, C., & Hort, J. (2013). The influence of sensory and packaging cues on both liking and emotional, abstract and functional conceptualisations. *Food Quality and Preference, 29*(2), 146–156.

Oppermann, A. K. L., de Graaf, C., Scholten, E., Stieger, M., & Piqueras-Fiszman, B. (2017). Comparison of Rate-All-That-Apply (RATA) and Descriptive sensory Analysis (DA) of model double emulsions with subtle perceptual differences. *Food Quality and Preference, 56*, 55–68.

Ramirez, G., Hough, G., & Contarini, A. (2001). Influence of Temperature and Light Exposure on Sensory Shelf-Life of a Commercial Sunflower Oil. *Journal of Food Quality, 24*(3), 195–204.

Ratinaud, P. (2014). IRaMuTeQ : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. In. France.

R Core Team (2018). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rios-Mera, J. D., Saldana, E., Cruzado-Bravo, M. L. M., Patinho, I., Selani, M. M., Valentin, D., et al. (2019). Reducing the sodium content without modifying the quality of beef burgers by adding micronized salt. *Food Res Int, 121*, 288–295.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Comparing Check-All and Forced-Choice Question Formats in Web Surveys. *Public Opinion Quarterly, 70*(1), 66–77.

Stampanoni, C. R. (1993). *The quantitative profiling technique. Perfumer Flavourist, 18*, 19–24.

Stone, H., Sidel, J. L., Oliver, S., Woolsey, A., & Singleton, R. C. (1974). Sensory evaluation by quantitative descriptive analysis. *Food Technology, 28*(11), 24–33.

Sudman, S., & Bradburn, M. B. (1982). *Asking questions.* San Francisco: Jossey-Bass.

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.

Tijssen, I. O. J. M., Zandstra, E. H., den Boer, A., & Jager, G. (2019). Taste matters most: Effects of package design on the dynamics of implicit and explicit product evaluations over repeated in-home consumption. *Food Quality and Preference, 72*, 126–135.

Valentin, D., Chollet, S., Lelièvre, M., & Abdi, H. (2012). Quick and dirty but still pretty good: A review of new descriptive methods in food science. *International Journal of Food Science & Technology, 47*(8), 1563–1578.

Varela, P., & Ares, G. (2012). Sensory profiling, the blurred line between sensory and consumer science. A review of novel methods for product characterization. *Food Research International, 48*(2), 893–908.

C. Stability of the provided descriptive sensory information

Article published in Food Quality and Preference:

An investigation of the stability of Free-Comment and Check-All-That-Apply in two consumer studies on red wines and milk chocolates

Benjamin Mahieu [a,*], Michel Visalli [a], Arnaud Thomas [b], Pascal Schlich [a,*]

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

[b] *SensoStat, Dijon, France*

Reference:

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

Food Quality and Preference 90 (2021) 104159

Contents lists available at ScienceDirect

Food Quality and Preference

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

An investigation of the stability of Free-Comment and Check-All-That-Apply in two consumer studies on red wines and milk chocolates

Benjamin Mahieu [a,*], Michel Visalli [a], Arnaud Thomas [b], Pascal Schlich [a,*]

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

ARTICLE INFO

Keywords:
Open-ended questions
Stability
Sensory method comparison
Consumer study

ABSTRACT

Free-Comment (FC), as a response to open-ended questions, enables a word-based sensory description and discrimination of sets of products. The stability of FC outputs has never been investigated and is the purpose of the present paper. Since Check-All-That-Apply (CATA) is the most popular method for the word-based sensory description of products with consumers, the stability of FC was compared to that of CATA performed on the same products. Four red wines and four milk chocolates were evaluated according to different sensory modalities by groups of consumers following either an FC or a CATA protocol. The stability of the product configurations and the product by descriptor associations were investigated. FC outputs were slightly more stable than CATA ones. Sixty consumers enable to guarantee medium stability, if not good, of FC and CATA outputs when the investigated product space is characterized by large differences between the products. The minimum number of consumers to obtain stable results was strongly dependent on the size of the differences between the products, which suggests that if *a priori* knowledge on the size of the differences between the investigated products is available, it must drive the decision of the number of consumers to include in the study rather than relying on an absolute rule. For both FC and CATA, the product configurations were more easily stable in terms of numbers of consumers than the product by descriptor associations. Investigating the stability of the product by descriptor associations *a posteriori* is recommended for future FC and CATA studies.

1. Introduction

Free-Comment (FC) (ten Kleij & Musters, 2003), as a response to open-ended questions, is a sensory method that enables collecting word-based sensory descriptions of a set of products without a predefined list of descriptors. For each evaluated product, consumers are asked to describe the product in their own words (Ares, Giménez, Barreiro, & Gámbaro, 2010; Hanaei, Cuvelier, & Sieffermann, 2015; Lahne, Trubek, & Pelchat, 2014; Luc, Lê, & Philippe, 2020; Mahieu, Visalli, Thomas, & Schlich, 2020; Symoneaux, Galmarini, & Mehinagic, 2012; ten Kleij & Musters, 2003). FC has already proven itself an efficient method in characterizing and discriminating sets of products both with consumers and experts (Lahne et al., 2014; Lawrence et al., 2013; ten Kleij & Musters, 2003) even out of the lab (Mahieu et al., 2020).

Check-All-That-Apply (CATA) (Adams, Williams, Lancaster, & Foley, 2007) is a sensory method based on a predefined list of descriptors that enables collecting word-based sensory descriptions of sets of products. For each evaluated product, consumers are asked to choose among a list of descriptors, those that apply to the product. CATA also has proven itself an efficient method for the characterization and discrimination of sets of products with consumers (Oppermann, de Graaf, Scholten, Stieger, & Piqueras-Fiszman, 2017; Valentin, Chollet, Lelièvre, & Abdi, 2012; Varela & Ares, 2012).

Probably because of the lack of tools for FC data analysis and ease of use of CATA, CATA is the most popular method for the word-based description of products with consumers. However, FC can provide better product discrimination as well as a richer characterization of the products as compared to CATA (Mahieu et al., 2020). Yet, while CATA has been suggested to provide stable outputs with a minimum of 60–80 consumers when differences between the products are large (Ares, Tárrega, Izquierdo, & Jaeger, 2014), the stability of the outputs provided by FC remains an open question.

* Corresponding authors.
E-mail addresses: benjamin.mahieu@inrae.fr (B. Mahieu), pascal.schlich@inrae.fr (P. Schlich).

https://doi.org/10.1016/j.foodqual.2020.104159
Received 17 July 2020; Received in revised form 17 November 2020; Accepted 20 December 2020
Available online 22 December 2020

In addition to the ability to characterize and discriminate the products, it is assumed that sensory methods should provide similar outputs across repeated experiments conducted in similar experimental settings. In consumer studies, it is also assumed that the larger the consumer panel, the more stable the outputs should be, but the more expensive the study is in terms of time and budget. For these reasons, having *a priori* knowledge of the number of consumers necessary to obtain stable outputs is important.

For consumer-oriented sensory methods, gathering a large number of different experiments conducted under similar experimental settings with different panel sizes is nearly impossible for practical limitations (Ares, Tárrega et al., 2014). Thus, the stability of the outputs is often evaluated internally, rather than externally, using bootstrap resampling of an actual panel that performed a study in the experimental settings under interest (Ares, Bruzzone et al., 2014; Ares, Tárrega et al., 2014; Blancher, Clavier, Egoroff, Duineveld, & Parcon, 2012; Cadena et al., 2014; Mammasse & Schlich, 2014; Vidal et al., 2014; Vidal, Tárrega, Antúnez, Ares, & Jaeger, 2015). This procedure enables to generate a large number of virtual panels of different sizes that simulate repeated experiments under similar experimental settings. The outputs obtained from the actual panel are considered as a benchmark to which those of the virtual panels are compared.

Depending on the sensory method under investigation, different aspects of the outputs are compared between the actual and the virtual panels. The product configurations between the actual and the virtual panels were compared in every aforementioned study using the RV coefficient (Escoufier, 1973; Robert & Escoufier, 1976). For word-based sensory methods, the descriptor configurations were also compared using the RV coefficient (Ares, Bruzzone et al., 2014; Ares, Tárrega et al., 2014; Vidal et al., 2015). However, the descriptor configurations are usually not interpreted for themselves but rather together with the product configurations to characterize the product space. Thus, investigating the stability of the product by descriptor associations rather than the stability of the descriptor configurations seems to be more in line with common practices.

To the best of our knowledge, in the context of consumer word-based sensory methods, no methodology has been proposed in the literature to compare the outputs of the product by descriptor associations of the actual and the virtual panels. The present paper proposed a methodology to do so and applied it on 10 datasets corresponding to the evaluation of red wines and milk chocolates on different sensory modalities by consumers using FC or CATA. The first objective was to investigate the number of consumers necessary to ensure the stability of FC outcomes. The second objective was to compare FC and CATA conducted in similar experimental settings on the stability of the outputs they provided.

2. Material and methods

2.1. Datasets

The information concerning the datasets used in this paper and provided across the material and methods section are summarized in Table 2.

All the data were collected using TimeSens® software (INRAE, Dijon, France).

2.1.1. First study: red wines

The datasets of this study are the same from Mahieu et al. (2020).

2.1.1.1. Participants. One-hundred and twenty consumers being 18 to 60 years old participated in this study. They were recruited from a population registered in the ChemoSens Platform's PanelSens database. This database has been declared to the relevant authority (Commission Nationale Informatique et Libertés—CNIL—n° d'autorisation 1148039). The consumers recruited were consumers of red wines at least once every two weeks and were allocated in two groups of 60 consumers. The two groups were balanced in terms of age repartition and gender and they were matched for consumption frequency. The first group performed an FC task while the second group performed a CATA task. Both FC and CATA were performed at home.

2.1.1.2. Products. Four commercialized French red wines from different terroirs were used. The four terroirs were Bordeaux, Beaujolais, Languedoc and Val de Loire.

2.1.1.3. FC task and datasets. For each red wine, the FC task was carried out by sensory modality in the following order: visual, olfactory, and gustatory. For each sensory modality, the following instructions were given to the consumers:

- Visual: "Describe the visual characteristics of the wine"
- Olfactory: "Describe the olfactory characteristics of the wine"
- Gustatory: "Describe the gustatory characteristics of the wine"

No particular restriction was given to the consumers on the manner of stating their descriptions.

The evaluations of the red wines using FC according to the three sensory modalities provided three distinct datasets named FC-Wine-Vis, FC-Wine-Olf, and FC-Wine-Gus.

2.1.1.4. CATA task and datasets. For each red wine, the CATA task was carried out by sensory modality in the following order: visual, olfactory, and gustatory. The gustatory description was presented in two steps to the consumers: they first evaluated the basic tastes and then the aromas. For each sensory modality, the following instruction was given to the consumers:

"Check in the subsequent list the words that apply to this wine".

The CATA lists of visual, olfactory, and gustatory descriptors were composed of 8, 10, and 19 descriptors respectively. The visual descriptors were the following: violet, opaque, dull, light red, bright, deep red, black, and transparent. The olfactory descriptors were the following: black fruit, roasted, red fruit, green vegetable, peppery/spicy, ripe fruit, animal, undergrowth, herbaceous, and woody. The gustatory descriptors were the following: alcohol, slight, astringent, bitter, concentrated, balanced, sweet, persistent, sour, red fruit, ripe fruit, green vegetable, black fruit, roasted, peppery/spicy, herbaceous, woody, undergrowth, and animal. These descriptors were selected according to the expertise of wine professionals, considering that they should be understandable by consumers, and were presented in a different randomized order for each consumer but with a constant order across evaluations for a given consumer.

The evaluations of the red wines using CATA according to the three sensory modalities provided three distinct datasets named CATA-Wine-Vis, CATA-Wine-Olf, and CATA-Wine-Gus.

2.1.2. Second study: milk chocolates

2.1.2.1. Participants. One-hundred and forty-seven consumers being 18 to 65 years old participated in this study. Seventy-seven of them were recruited from a population registered in the ChemoSens Platform's PanelSens database and performed an FC task at home. The remaining seventy consumers were employees of the Barry Callebaut© Company (not implied in sensory and consumer research) and performed a CATA task in a dedicated room at the Barry Callebaut© Company. The consumers recruited were consumers of milk chocolates at least once every two weeks and were not involved in the first study. The two groups were balanced in terms of age repartition and gender.

2.1.2.2. Products. Four milk chocolate with different recipes were used: a standard Belgian milk chocolate, a Swiss milk chocolate, a milk

compound chocolate, and a protein base milk chocolate.

2.1.2.3. FC task and datasets. For each milk chocolate, the FC task was carried out by sensory modality in the following order: texture and flavor in the mouth. For each sensory modality, the following instructions were given to the consumers:

- Mouth texture: "Describe the mouth texture characteristics of the chocolate"
- Mouth flavor: "Describe the mouth flavor characteristics of the chocolate"

No particular restriction was given to the consumers on the manner of stating their descriptions.

The evaluations of the milk chocolates using FC according to the two sensory modalities provided two distinct datasets named FC-Choc-Tex and FC-Choc-Fla.

2.1.2.4. CATA task and datasets. For each milk chocolate, the CATA task was carried out by sensory modality in the following order: texture and flavor in the mouth. For each sensory modality, the following instruction was given to the consumers:

"Check in the subsequent list the words that apply to this chocolate".

The CATA lists of mouth texture and mouth flavor descriptors were composed of 8 and 6 descriptors respectively. The mouth texture descriptors were the following: hard, soft, sticky, melting, coarse, fatty, creamy texture, and mouthcoating. The mouth flavor descriptors were the following: sweet, bitter, cocoa, caramel, cereal, and milky. These descriptors were selected according to the expertise of Barry Callebaut© and were presented in a different randomized order for each consumer but with a constant order across evaluations for a given consumer.

The evaluations of the milk chocolates using CATA according to the two sensory modalities provided two distinct datasets named CATA-Choc-Tex and CATA-Choc-Fla.

2.2. Data treatment

2.2.1. FC data treatment

All the FC data treatments were performed using R 3.5.1 (R Core Team. (2018), 2018). The lexicon provided with IRaMuTeQ© (Ratinaud, 2014) software was used for lemmatization and part-of-speech tagging. The FC datasets were treated separately with the method described in Mahieu et al. (2020) and summarized thereafter.

The descriptions were first cleaned, lemmatized, and filtered. Then, the words with similar meanings were grouped into latent-words relying on a chi-square-distance-based ascendant hierarchical classification.

Among all the words and latent words, only those mentioned by at least 5% of the panel for at least one product were retained for further analysis and called descriptors thereafter. The FC lists of descriptors were composed of 8 to 20 descriptors.

The number of times each descriptor was cited for each product was computed at the panel level. Then, the corresponding contingency table containing the citation counts of each descriptor for each product was built.

2.2.2. CATA data treatment

The CATA datasets were treated separately and identically. The number of times each descriptor was checked for each product was computed at the panel level. Then, the corresponding contingency table containing the citation counts of each descriptor for each product was built.

2.3. Data analyses

All analyses were performed using R 3.5.1 (R Core Team. (2018), 2018).

2.3.1. Similarity of FC and CATA outputs

For each pair product/sensory-modality, the RV coefficient (Escoufier, 1973; Robert & Escoufier, 1976) between the configuration provided by FC and CATA was computed.

2.3.2. Size of the differences between the products

For each contingency table, the following quantity (called Cramér's Phi coefficient in the present paper) was computed as originally proposed by (Cramér, 1946):

$$\phi_C = \frac{\phi^2}{\min(r-1, c-1)}$$

with ϕ^2 the phi-square index of the contingency table, r the number of rows of the contingency table, and c the number of columns of the contingency table. The phi-square index is equal to the sum of the eigenvalues associated with the Correspondence Analysis (CA) of the contingency table. The minimum between $r-1$ and $c-1$ is the total number of axes of this CA. Like the phi-square index itself, the Cramér's Phi coefficient is a measure of the intensity of the dependence between rows and columns of contingency tables. Intuitively, Cramér's Phi coefficient represents the average dependence captured by one CA axis. The benefit of the Cramér's Phi coefficient over the phi-square index is that it provides a measure that is comparable when contingency tables are of different sizes. Cramér's Phi coefficient ranges between 0 (independence) and 1 (full dependence, which corresponds to a diagonal contingency table).

In the case of word-based sensory methods, the closer to 1 the Cramér's Phi coefficient, the more dependence between products and descriptors exists in the contingency table, and thus the more different the products are. The size of the differences between the products on a given sensory modality is estimated thanks to the Cramér's Phi coefficient in both CATA and FC. The Cramér's Phi coefficients were compared from one dataset to another to obtain a relative ranking of the datasets in terms of size of differences between the products. For an absolute interpretation, one can refer for example to Cohen (1988).

2.3.3. Stability of the outputs

For all computations described in this section, the configurations were obtained by CA of the contingency tables. Principal coordinates of the products and contribution coordinates of the descriptors were used (Castura, Antúnez, Giménez, & Ares, 2016; Greenacre, 2013).

The stability of the descriptor configurations was not investigated (Ares, Bruzzone et al., 2014; Ares, Tárrega et al., 2014; Vidal et al., 2015) because they are usually not interpreted for themselves but rather as help for interpretation to understand the product configurations. In this sense, the stability of the joint product by descriptor configurations and of the product by descriptor significant associations were investigated instead. The choice to keep two indicators (joint product by descriptor configurations and product by descriptor significant associations) that seem similar is deliberate. The joint product by descriptor configurations corresponds to the product by descriptor insights one would draw from reading the map and/or the space resulting from the CA of the contingency table. By nature, this reading is subjective and approximate but has the benefit of being nuanced. The product by descriptor significant associations are the black and white version of the joint product by descriptor configurations and corresponds to the product by descriptor insights one would draw from reading the tables as presented Mahieu et al. (2020). By their statistical-based nature, the product by descriptor significant associations are objective but have the drawback of being threshold-dependent and binary.

2.3.3.1. Bootstrap resampling procedure. For each dataset, different sizes of virtual panels were considered ranging from 10 to the size of the actual panel, increasing with a step of 10. For each size, 1000 virtual panels were constituted. Each virtual panel was constituted by randomly drawing subjects from the actual panel with replacement. The outputs obtained from the actual panel were considered as a benchmark to which the outputs of the virtual panels were compared.

2.3.3.2. Product configurations. The product configurations, i.e. the relative position of the products in relation to each other in the sensory space, were compared by computing the RV coefficient (Escoufier, 1973; Robert & Escoufier, 1976) in the full space between the product configurations of the actual and the virtual panels.

2.3.3.3. Joint product by descriptor configurations. To compare the joint product by descriptor configurations, i.e. the position of each product in relation to the descriptor configuration in the sensory space, the scalar products in the full space between each product vector and each descriptor vector were computed for both the actual and the virtual panels. Then, these scalar products were vectorized and the Pearson correlation coefficient was computed between the vectorized vector of scalar products of the actual panel and those of the virtual panels.

2.3.3.4. Product by descriptor significant associations. Fisher's exact tests per cell with a one-sided greater alternative hypothesis were conducted on each contingency table. The tests were considered significant at the α-risk of 5%. These tests represent the binary statistical-based relations between each product with each descriptor.

To measure the similarity between the outputs of the tests obtained in the actual panel and each virtual panel, the Phi correlation coefficient was computed. The Phi correlation coefficient is defined as follows:

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

with "a" the number of tests that were significant in both the actual panel and the virtual panel, "b" the number of tests that were significant in the actual panel but not in the virtual panel, "c" the number of tests that were not significant in the actual panel but were in the virtual panel and "d" the number of tests that were not significant in both the actual panel and the virtual panel.

The Phi correlation coefficient is a measure of the correlation between two binary variables. It ranges between −1 and 1. A value of 0 indicates that the two variables are uncorrelated. In our case, the closer to 1 the Phi correlation coefficient, the more similar the product by descriptor significant associations were between the actual and the virtual panels.

2.3.3.5. Stability of outcomes. The reading grid was the same for all the coefficients. The stability was considered good when no more than 5% of the coefficients were below 0.80. The stability was considered poor when more than 5% of the coefficients were below 0.50. When the stability was neither good nor poor, it was considered medium. These thresholds were selected according to a common absolute value (considering that in an ideal world they should be equal to one). It was necessary to achieve an objective reading of the results. They were the same for the three correlation coefficients to allow for a relative comparison in terms of stability of the three aspects of the outputs investigated since each coefficient is comparable to the others. The proposed thresholds do not intend to become "gold standards". Other thresholds might have been considered and might be interesting in applications.

To compare the 5% quantile of the distributions of the correlation coefficients to the different thresholds rather than the mean of these distributions (Ares, Bruzzone et al., 2014; Ares, Tárrega et al., 2014; Blancher et al., 2012; Cadena et al., 2014; Vidal et al., 2014) is more in line to what a virtual panel drawn from the bootstrap resampling of the actual panel represents. Indeed, under the hypothesis where such a virtual panel represents a new study conducted in similar experimental settings, similar outputs to those of the actual panel considered as a benchmark are expected from this virtual panel. Thus, high correlation coefficients between the outputs of the actual and the virtual panel are expected. Extended to a large number of virtual panels, this line of reasoning still holds, and thus considering the entire distribution rather than its mean is more in line with the bootstrap hypothesis made and with what a virtual panel represents.

3. Results

3.1. Similarity of FC and CATA outputs

Overall, Table 1 shows that the RV coefficients between FC and CATA configurations are high, which indicates that they provided similar product configurations.

On the detailed characterization provided by FC and CATA about the products, the reader can refer to Mahieu et al. (2020) concerning the red wines. For the milk chocolates, the characterization provided by FC and CATA were overall similar: the same sensory dimensions discriminated the products.

3.2. Size of the differences between the products

Table 2 summarizes the characteristics and the measures of the size of the differences between the products for each dataset. For FC Cramér's Phi coefficient ranged between 0.05 (Wine-Olf) and 0.17 (Choc-Tex). For CATA Cramér's Phi coefficient ranged between 0.02 (Wine-Olf and Wine-Gus) and 0.20 (Choc-Tex). This suggests that the size of the differences between the products differed from one product type to another and from one sensory modality to another. For both FC and CATA, Cramér's Phi coefficients were lower for the red wines than for the milk chocolates suggesting that the size of the differences was lower between the red wines than between the milk chocolates.

3.3. Stability of the outputs

3.3.1. Product configurations

Fig. 1 shows that good stability of the product configurations was reached for Wine-Gus, Choc-Tex, and Choc-Fla with the same minimum number of consumers with FC and CATA, respectively with 10, 10, and 20 consumers. For Wine-Vis and Wine-Olf, good stability was reached with FC with fewer consumers as compared to CATA (20 vs. 40 for Wine-Vis, 30 vs. no good stability for Wine-Olf).

Overall, the average stability of the product configurations for a given size of virtual panels and a given pair product / sensory-modality was almost the same between FC and CATA but the minimum number of consumers required to obtain good stability of the product configurations whatever the dataset was 30 for FC, and 40 for CATA (except for CATA-Wine-Olf, which never reached good stability) and good stability was reached in more datasets with FC than with CATA (5 vs. 4). For both FC and CATA, the stability of product configurations was higher for the

Table 1
RV coefficients between FC and CATA configurations for each pair product / sensory-modality.

Product type	Sensory modality	RV coefficient between FC and CATA configurations
Red wine	Visual	0.90
Red wine	Olfactory	0.84
Red wine	Gustatory	0.86
Milk chocolate	Mouth texture	0.93
Milk chocolate	Mouth flavor	0.98

Table 2
Characteristics and measure of the size of the differences between the products for each dataset.

Dataset	Product type	Sensory modality	Sensory method	Number of products	Number of subjects	Number of descriptors	Measure of the size of the differences between the products (ϕ_C)
FC-Wine-Vis	Red wine	Visual	FC	4	60	12	0.06
CATA-Wine-Vis	Red wine	Visual	CATA	4	60	8	0.03
FC-Wine-Olf	Red wine	Olfactory	FC	4	60	14	0.05
CATA-Wine-Olf	Red wine	Olfactory	CATA	4	60	10	0.02
FC-Wine-Gus	Red wine	Gustatory	FC	4	60	20	0.07
CATA-Wine-Gus	Red wine	Gustatory	CATA	4	60	19	0.02
FC-Choc-Tex	Milk chocolate	Mouth texture	FC	4	77	10	0.17
CATA-Choc-Tex	Milk chocolate	Mouth texture	CATA	4	70	8	0.20
FC-Choc-Fla	Milk chocolate	Mouth flavor	FC	4	77	8	0.13
CATA-Choc-Fla	Milk chocolate	Mouth flavor	CATA	4	70	7	0.14

chocolate datasets, for which the size of the product differences was higher.

3.3.2. Joint product by descriptor configurations

Fig. 2 shows that whatever the method, good stability of the joint product by descriptor configurations was not reached for Wine-Olf with the actual number of consumers. For Wine-Vis and Wine-Gus, good stability was reached with FC with fewer consumers compared to CATA (40 vs. 50 for Wine-Vis, 60 vs. no good stability for Wine-Gus). For Choc-Tex and Choc-Fla, good stability was reached with FC with more consumers compared to CATA (20 vs. 10 for Choc-Tex, 30 vs. 20 for Choc-Fla).

Overall, the minimum number of consumers required to obtain good stability of the joint product by descriptor configurations whatever the dataset was more than 60 consumers for both FC and CATA but the average stability for a given pair product / sensory-modality with 60 consumers and more was slightly higher with FC than with CATA for some datasets (Wine-Olf and Wine-Gus) and stability was reached in more datasets with FC than with CATA (4 vs. 3). For both FC and CATA, the stability of the joint product by descriptor configurations increased with the size of the product differences of the datasets. For both FC and CATA, the stability of joint product by descriptor configurations was higher for the chocolate datasets, for which the size of the product differences was higher.

3.3.3. Product by descriptor significant associations

Fig. 3 shows that whatever the method, good stability of the product by descriptor significant associations was not reached with the actual number of consumers for all datasets and the stability was poor for the red wines datasets with the actual number of consumers. Medium stability of the product by descriptor significant associations was reached for Choc-Tex with 30 consumers for FC and 20 consumers for CATA, and for Choc-Fla with 30 consumers for FC and 50 consumers for CATA.

Overall, the minimum number of consumers required to obtain at least moderately stable product by descriptor significant associations whatever the dataset was more than 60 consumers for both FC and CATA but the average stability for a given pair product / sensory-modality was higher with FC than with CATA with 60 consumers and more for all datasets except Choc-Text. For both FC and CATA, the stability of product by descriptor significant associations was higher for the chocolate datasets, for which the size of the product differences was higher.

4. Discussion

4.1. The stability of the outputs provided by FC and CATA

Results showed relatively stable FC outputs, at least as stable as CATA ones if not more. FC outputs reached good stability in more datasets than CATA ones regarding product configurations and joint product by descriptor configurations. Further, the average stability of FC outputs was always larger than or equal to CATA ones for the three aspects of the outputs investigated in this study when a given pair product / sensory-modality with 60 consumers and more was considered. These results suggest that FC outputs are on the same level of stability than CATA ones, at least when FC and CATA are performed by sensory modality. Future studies need to be conducted to confirm or refute these results when FC and CATA are performed with a single overall characterization of each product (not by sensory modality).

Two experimental points should be outlined. First, the consumers who performed the chocolate CATA task might be more knowledgeable about chocolate than if they were naïve consumers. Thus, the CATA descriptions might have been more consensual, which might have resulted in higher stability of the outputs. Therefore, the stability of CATA outputs might have been overestimated in the chocolate study. Second, some descriptors of the CATA list in the wine study may be considered reasonably technical (e.g. animal, roasted, etc.). This may have impeded the agreement of consumers on CATA descriptions, which may have resulted in lesser stability of the outputs. However, some of these “technical descriptors” were mentioned during the FC task (Mahieu et al., 2020), which suggests that they were meaningful to consumers. They were however mentioned less frequently in FC as compared to CATA, but so were common descriptors shared by FC and CATA (Mahieu et al., 2020). Indeed, the CATA task encourages consumers to check the proposed descriptors (Callegaro, Murakami, Tepman, & Henderson, 2015; Kim, Hopkinson, van Hout, & Lee, 2017; Krosnick, 1999). This suggests that this difference in citation frequency is due to the task and not to the potential “technical” aspect of the descriptors.

Not surprisingly, for both FC and CATA, the stability of the product configurations increased with the size of the virtual panel and with the size of the differences between the products. The minimum number of subjects to obtain stable product configurations was of the same order of magnitude than previously reported for CATA, Rate-All-That-Apply, Projective Mapping, Sorting, and Polarized Sensory Positioning (Ares, Bruzzone et al., 2014; Ares, Tárrega et al., 2014; Blancher et al., 2012; Cadena et al., 2014; Vidal et al., 2015).

The overall level of stability was more impacted by the size of

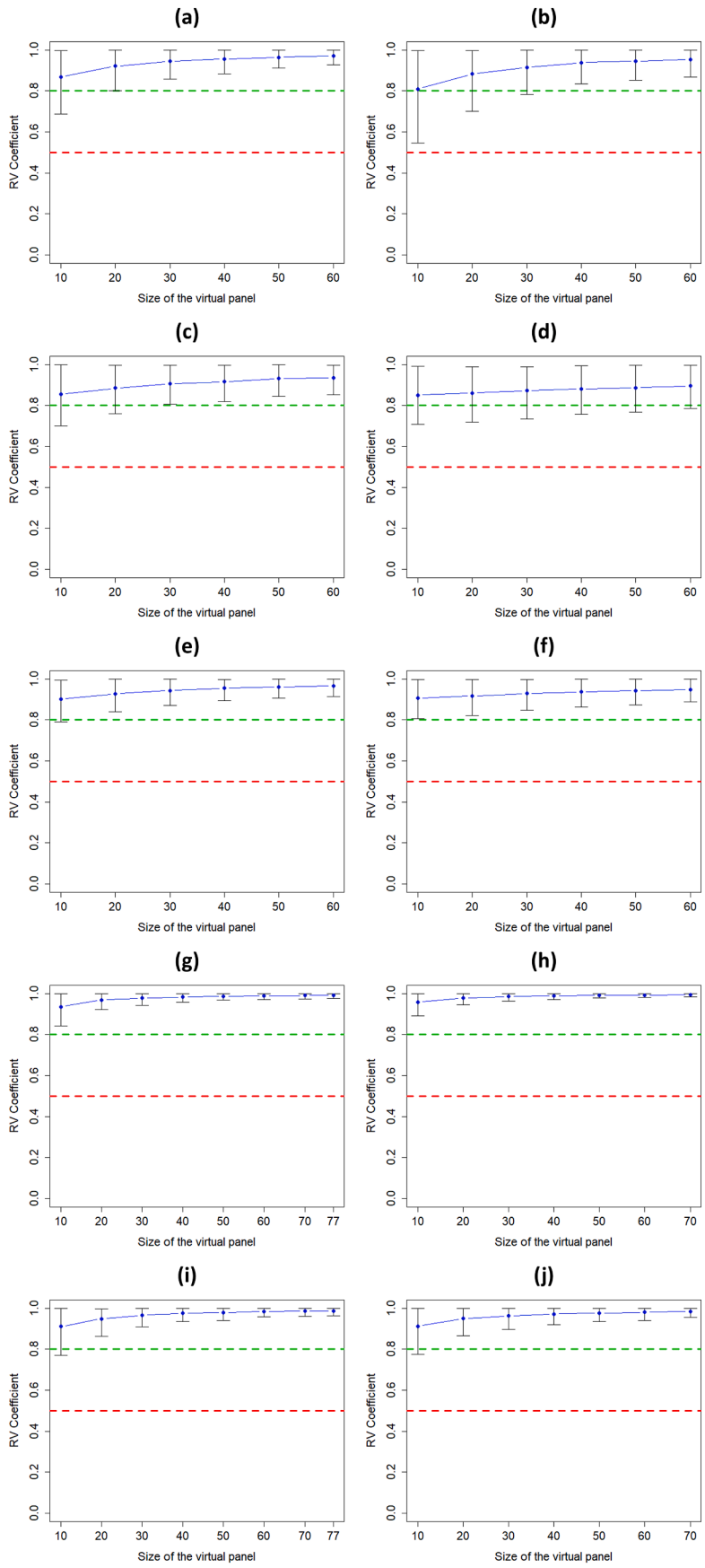


Fig. 1. Mean of the distribution of the RV coefficients between the actual and the virtual product configurations as a function of the virtual panel size for (a) FC-Wine-Vis, (b) CATA-Wine-Vis, (c) FC-Wine-Olf, (d) CATA-Wine-Olf, (e) FC-Wine-Gus, (f) CATA-Wine-Gus, (g) FC-Choc-Tex, (h) CATA-Choc-Tex, (i) FC-Choc-Fla and (j) CATA-Choc-Fla. Dashed lines indicates 0.80 (green) and 0.50 (red). Error bars show the 0.05 and 1 quantiles of the distributions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

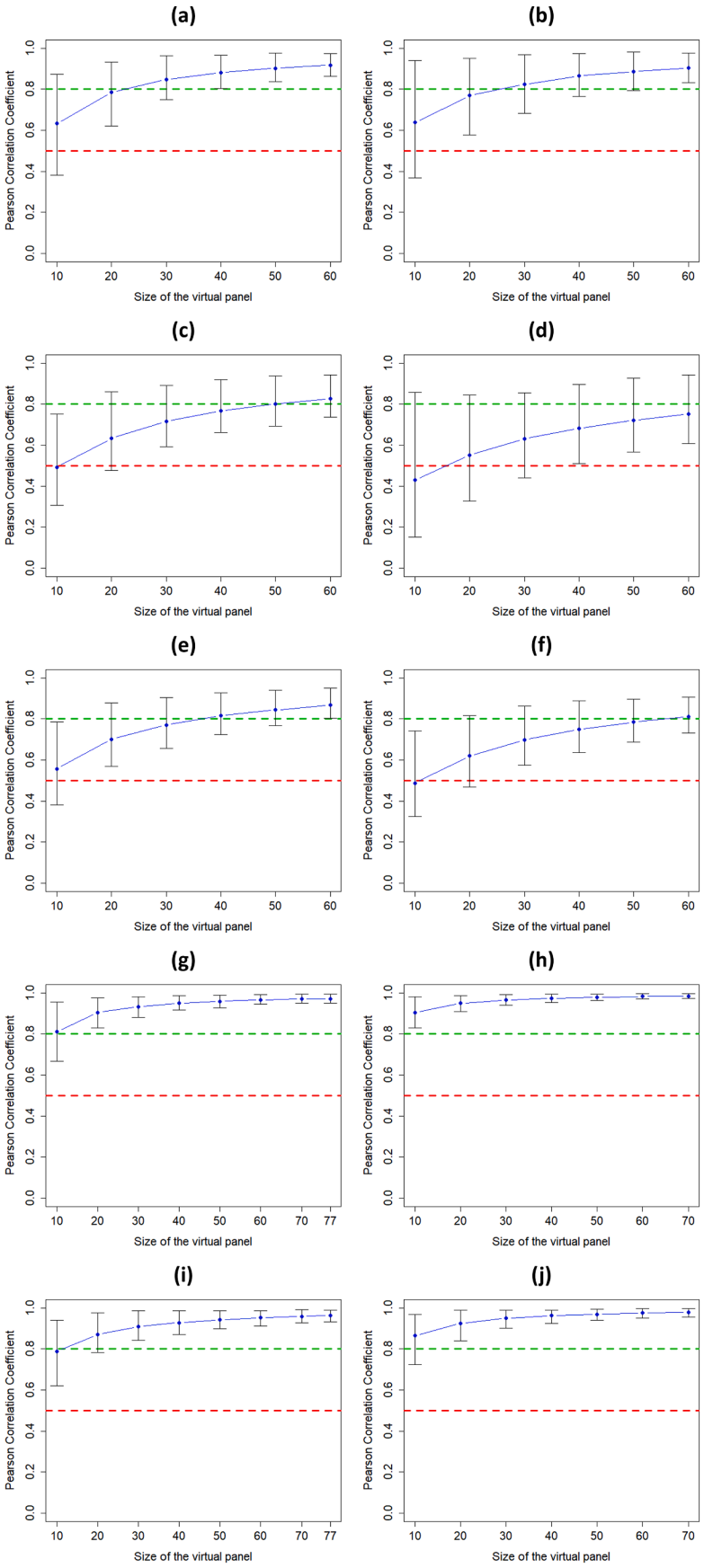


Fig. 2. Mean of the distribution of the Pearson correlation coefficients between the actual and the virtual joint product by descriptor configurations as a function of the virtual panel size for (a) FC-Wine-Vis, (b) CATA-Wine-Vis, (c) FC-Wine-Olf, (d) CATA-Wine-Olf, (e) FC-Wine-Gus, (f) CATA-Wine-Gus, (g) FC-Choc-Tex, (h) CATA-Choc-Tex, (i) FC-Choc-Fla and (j) CATA-Choc-Fla. Dashed lines indicates 0.80 (green) and 0.50 (red). Error bars show the 0.05 and 1 quantiles of the distributions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

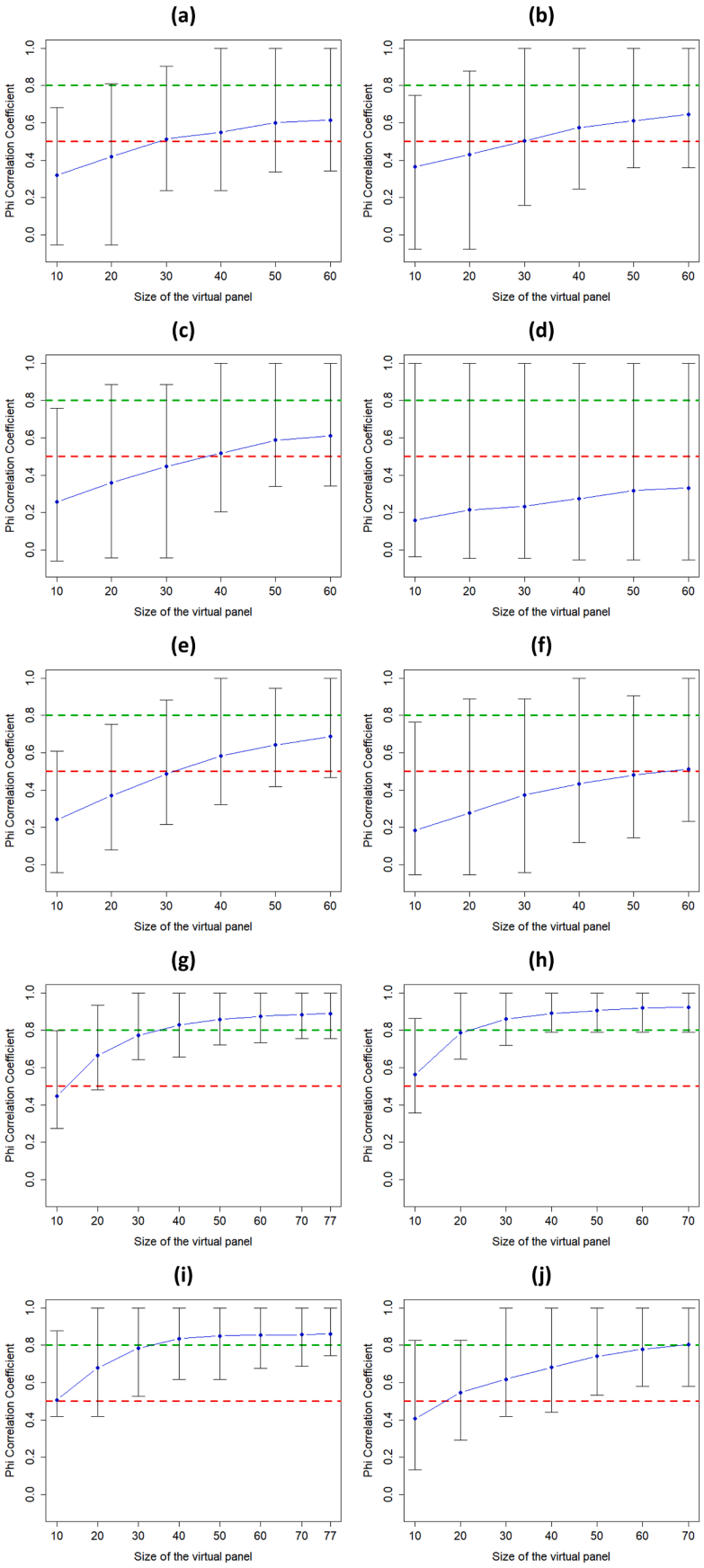


Fig. 3. Mean of the distribution of the Phi correlation coefficients between the actual and the virtual Fisher's exact tests per cell (α = 5%) outputs as a function of the virtual panel size for (a) FC-Wine-Vis, (b) CATA-Wine-Vis, (c) FC-Wine-Olf, (d) CATA-Wine-Olf, (e) FC-Wine-Gus, (f) CATA-Wine-Gus, (g) FC-Choc-Tex, (h) CATA-Choc-Tex, (i) FC-Choc-Fla and (j) CATA-Choc-Fla. Dashed lines indicates 0.80 (green) and 0.50 (red). Error bars show the 0.05 and 1 quantiles of the distributions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

product differences than by the method used (FC versus CATA). These results are in line with some previously reported studies (Ares, Bruzzone et al., 2014; Ares, Tárrega et al., 2014; Blancher et al., 2012; Mammasse & Schlich, 2014; Vidal et al., 2015), even with sensory descriptive analysis (Gacula & Rutenbeck, 2006; Heymann, Machado, Torri, & Robinson, 2012; Silva, Minim, Silva, & Minim, 2014). This effect of the size of product differences affected the stability of both FC and CATA in the same direction and with the same magnitude.

For both FC and CATA, the product configurations were more stable than the joint product by descriptor configurations, themselves being more stable than the product by descriptor significant associations. This suggests that the more an aspect of the outputs is demanding, the less it is stable. The product configurations are relatively stable because they are driven by intrinsic differences between the products and do not depend on how these intrinsic differences are transcribed and/or verbalized. This is supported by several studies that compared two or more consumer sensory methods and observed that they provided similar product configurations (Ares, Bruzzone et al., 2014; Fleming, Ziegler, & Hayes, 2015; Oppermann et al., 2017; Reinbach, Giacalone, Ribeiro, Bredie, & Frøst, 2014). The joint product by descriptor configurations are less stable than the product configuration because identifying differences is easier than explicitly verbalizing them. However, the joint product by descriptor configurations are still relatively stable because the big picture of each joint product by descriptor configuration is likely to be recovered across repeated experiments. The product by descriptor significant associations are at best moderately stable because they require the intrinsic product differences to be verbalized significantly with the same descriptors across repeated experiments, which is the most demanding aspect of the outputs.

4.2. Recommendations

When the investigated product space is characterized by large differences between the products, 60 consumers enable to guarantee at least a medium stability of FC and CATA outputs, which is in line with previous results concerning CATA (Ares, Tárrega et al., 2014). When differences between the products are more subtle, 60 consumers enable to guarantee at least a medium stability of the product configurations and the joint product by descriptor configurations for both FC and CATA but do not guarantee stable product by descriptor significant associations. Future studies need to be conducted to investigate the number of consumers necessary to obtain good stability of the product by descriptor significant associations when working with products that have subtle differences between them.

The previous recommendations are worthy of being nuanced by the fact that the stability of the outputs highly depends on the size of the differences between the products. Thus, these recommendations should be considered as an order of magnitude rather than an absolute rule. If the practitioner has *a priori* knowledge of the size of the differences between the products investigated, this information must be the principal driver to decide the number of consumers to include in the study. Practically, this *a priori* knowledge can arise from the relative comparison in terms of product differences of the product space investigated to product spaces previously investigated for which the stability of the outputs could have been investigated *a posteriori*.

Finally, like several authors recommended for the product configurations (Ares, Tárrega et al., 2014; Blancher et al., 2012; Vidal et al., 2014), investigating *a posteriori* the stability of the joint product by descriptor configurations and of the product by descriptor significant associations is recommended to determine the degree of confidence one should have in the product by descriptor insights obtained from the study.

5. Conclusion

FC outputs were slightly more stable than CATA ones. When the product space investigated is characterized by large differences between the products, 60 consumers enable to guarantee medium stability, if not good, of FC and CATA outputs. The minimum number of consumers to obtain stable results was strongly dependent on the size of the differences between the products, which suggests that if *a priori* knowledge on the size of the differences between the products investigated is available, it must drive the decision of the number of consumers to include in the study rather than an absolute rule. For both FC and CATA, the sensory spaces obtained from Correspondence Analysis were more stable than the product by descriptor significant associations obtained from Fisher's exact tests per cell. Among sensory spaces, the product configurations were more stable than the joint product by descriptor configurations. Finally, the stability of joint product by descriptor configurations and product by descriptor significant associations are recommended to be investigated *a posteriori* in the same manner that the stability of product configurations is.

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 Robert et Marcel®, Sicarex®, and Barry Callebaut® for providing their products.

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.

Ares, G., Bruzzone, F., Vidal, L., Cadena, R. S., Giménez, A., Pineau, B., … Jaeger, S. R. (2014). Evaluation of a rating-based variant of check-all-that-apply questions: Rate-all-that-apply (RATA). *Food Quality and Preference, 36*, 87–95.

Ares, G., Giménez, A., Barreiro, C., & Gámbaro, A. (2010). Use of an open-ended question to identify drivers of liking of milk desserts. Comparison with preference mapping techniques. *Food Quality and Preference, 21*(3), 286–294.

Ares, G., Tárrega, A., Izquierdo, L., & Jaeger, S. R. (2014). Investigation of the number of consumers necessary to obtain stable sample and descriptor configurations from check-all-that-apply (CATA) questions. *Food Quality and Preference, 31*, 135–141.

Blancher, G., Clavier, B., Egoroff, C., Duineveld, K., & Parcon, J. (2012). A method to investigate the stability of a sorting map. *Food Quality and Preference, 23*(1), 36–43.

Cadena, R. S., Caimi, D., Jaunarena, I., Lorenzo, I., Vidal, L., Ares, G., … Giménez, A. (2014). Comparison of rapid sensory characterization methodologies for the development of functional yogurts. *Food Research International, 64*, 446–455.

Callegaro, M., Murakami, M. H., Tepman, Z., & Henderson, V. (2015). Yes-no answers versus check-all in self-administered modes. *International Journal of Market Research, 57*, 203–223.

Castura, J. C., Antúnez, L., Giménez, A., & Ares, G. (2016). Temporal Check-All-That-Apply (TCATA): A novel dynamic method for characterizing products. *Food Quality and Preference, 47*, 79–90.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cramér, H. (1946). Chapter 21. The two-dimensional case. In P. U. Press, Mathematical Methods of Statistics.

Escoufier, Y. (1973). Le Traitement des Variables Vectorielles. *Biometrics, 29*(4), 751. https://doi.org/10.2307/2529140

Fleming, E. E., Ziegler, G. R., & Hayes, J. E. (2015). Check-all-that-apply (CATA), sorting, and polarized sensory positioning (PSP) with astringent stimuli. *Food Quality and Preference, 45*, 41–49.

Gacula, M., & Rutenbeck, S. (2006). Sample size in consumer test and descriptive analysis. *Journal of Sensory Studies, 21*(2), 129–145.

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

Hanaei, F., Cuvelier, G., & Sieffermann, J. M. (2015). Consumer texture descriptions of a set of processed cheese. *Food Quality and Preference, 40*, 316–325.

Heymann, H., Machado, B., Torri, L., & Robinson, A. L. (2012). How many judges should one use for sensory descriptive analysis? Journal of Sensory Studies, 27(2), 111-122.

Kim, I.-A., Hopkinson, A., van Hout, D., & Lee, H.-S. (2017). A novel two-step rating-based 'double-faced applicability' test. Part 1: Its performance in sample discrimination in comparison to simple one-step applicability rating. *Food Quality and Preference, 56*, 189–200.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537–567.

Lahne, J., Trubek, A. B., & Pelchat, M. L. (2014). Consumer sensory perception of cheese depends on context: A study using comment analysis and linear mixed models. *Food Quality and Preference, 32*, 184–197.

Lawrence, G., Symoneaux, R., Maitre, I., Brossaud, F., Maestrojuan, M., & Mehinagic, E. (2013). Using the free comments method for sensory characterisation of Cabernet Franc wines: Comparison with classical profiling in a professional context. *Food Quality and Preference, 30*, 145–155.

Luc, A., Lê, S., & Philippe, M. (2020). Nudging consumers for relevant data using Free JAR profiling: An application to product development. *Food Quality and Preference, 79*, 103751.

Mahieu, B., Visalli, M., Thomas, A., & Schlich, P. (2020). Free-comment outperformed check-all-that-apply in the sensory characterisation of wines with consumers at home. *Food Quality and Preference, 84*, 103937.

Mammasse, N., & Schlich, P. (2014). Adequate number of consumers in a liking test. Insights from resampling in seven studies. *Food Quality and Preference, 31*, 124–128.

Oppermann, A. K. L., de Graaf, C., Scholten, E., Stieger, M., & Piqueras-Fiszman, B. (2017). Comparison of Rate-All-That-Apply (RATA) and Descriptive sensory Analysis (DA) of model double emulsions with subtle perceptual differences. *Food Quality and Preference, 56*, 55–68.

R Core Team. (2018). R: A language and environment for statistical computing. In. Vienna, Austria: R Foundation for Statistical Computing.

Ratinaud, P. (2014). IRaMuTeQ : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. In. France.

Reinbach, H. C., Giacalone, D., Ribeiro, L. M., Bredie, W. L. P., & Frøst, M. B. (2014). Comparison of three sensory profiling methods based on consumer perception: CATA, CATA with intensity and Napping®. *Food Quality and Preference, 32*, 160–166.

Robert, P., & Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV- coefficient. *Applied Statistics, 25*, 257.

Silva, R.d. C. D. S. N.d., Minim, V. P. R., Silva, A. N.d., & Minim, L. A. (2014). Number of judges necessary for descriptive sensory tests. *Food Quality and Preference, 31*, 22–27.

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.

Valentin, D., Chollet, S., Lelièvre, M., & Abdi, H. (2012). Quick and dirty but still pretty good: a review of new descriptive methods in food science. International Journal of Food Science & Technology, 47(8), 1563-1578.

Varela, P., & Ares, G. (2012). Sensory profiling, the blurred line between sensory and consumer science. A review of novel methods for product characterization. Food Research International, 48(2), 893-908.

Vidal, L., Cadena, R. S., Antúnez, L., Giménez, A., Varela, P., & Ares, G. (2014). Stability of sample configurations from projective mapping: How many consumers are necessary? *Food Quality and Preference, 34*, 79–87.

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.

Chapter V:
Extensions of Free-Comment

A. Context and contents

The previous chapter confirmed that Free-Comment (FC) appears as a well-performing method for descriptive sensory analysis with consumers. Based on this observation considered together with the benefits of FC, it would be a pity to restrict the use of FC to the sole static sensory description of the products while other typical situations and problems occur in sensory analysis as shown in chapter I. This chapter proposes to remedy this limitation by proposing extensions of FC that can deal with temporal sensory analysis, driver of liking identification and ideal product characterization.

Section B tackles temporal sensory analysis by proposing the Free-Comment Attack-Evolution-Finish (FC-AEF) method following the Attack-Evolution-Finish (AEF) one. With FC-AEF, consumers report their perception at the three periods (Attack, Evolution and Finish) using a FC description for each period, rather than by selecting sensory descriptors among a pre-established list as in AEF. Two strategies are proposed to analyze FC-AEF data depending on the aims of the study. The first strategy consists in comparing the products at a given period (Figures 3 and 4 of section B). The second strategy consists in comparing the periods of a given product (Figures 5 and 6 of section B). An application of FC-AEF with 63 consumers evaluating five dark chocolates at home was conducted. Both strategies of analysis provided insightful and sensible temporal descriptive sensory information on the dark chocolates. This demonstrates that FC-AEF can provide temporal descriptive sensory information of a set of products without a pre-established list of sensory descriptors. FC-AEF thus opens new perspectives for temporal sensory analysis that is currently usually performed based on a necessary limited pre-established list of sensory descriptors as in Temporal Dominance of Sensations and Temporal-Check-All-That-Apply.

Section C tackles drivers of liking identification and ideal product characterization by proposing the Ideal-Free-Comment (IFC) method paired with

liking scoring. Three types of data are gathered in this method: the FC descriptions of the products under interest, the liking scores of the products under interest and the FC descriptions of the ideal product. Three strategies of analysis are proposed with these data. The first strategy consists in regressing liking scores of the actual products on the corresponding FC descriptions using a mixed linear model. This enables identifying positive and negative drivers of liking (Figure 1 of section C). The second strategy consists in estimating the proportion of citations of each sensory descriptor for the ideal product and testing them against the corresponding proportions for the pool of actual products. This enables characterizing the ideal product and its differences from the actual products (Figure 2 of section C). The third strategy consists in projecting the ideal product and the mean liking scores of the actual products in the sensory space depicted by multiple-response Correspondence Analysis of the characterization of the actual products. This enables to locate the ideal product relatively to the actual products and their liking scores (Figure 3 of section C). An application of IFC paired with liking scoring in a large study involving 483 consumers purchasing and evaluating cooked hams from a list of 30 hams representative of the French market was conducted. The number of hams evaluated by each consumer ranged between 1 and 14 (mean = 5.71, sd = 2.47) resulting in a total of 2758 evaluations. Each strategy of analysis provided insightful and sensible sensory information. The identified positive and negative drivers of liking made sense and the ideal product characterization was consistent with them and with liking scores of the actual products. This demonstrates that IFC paired with liking scoring can provide relevant information to understand preferences with no use of a pre-established list of sensory descriptors. IFC paired with liking thus opens new perspectives for hedonic optimization of products and understanding preferences without biasing consumers towards any sensory descriptor, and most importantly, by limiting chances of missing key information, unlike usual methods that are list-based.

B. Temporal sensory analysis: Free-Comment Attack-Evolution-Finish

Article published in Food Quality and Preference:

Using Free-Comment with consumers to obtain temporal sensory descriptions of products

Benjamin Mahieu[a,*], Michel Visalli[a], Arnaud Thomas[b], Pascal Schlich[a]

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

[b] *SensoStat, Dijon, France*

Reference:

Mahieu, B., Visalli, M., Thomas, A., & Schlich, P. (2020). Using Free-Comment with consumers to obtain temporal sensory descriptions of products. Food Quality and Preference, 86.

Food Quality and Preference 86 (2020) 104008

Contents lists available at ScienceDirect

Food Quality and Preference

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

Using Free-Comment with consumers to obtain temporal sensory descriptions of products

Benjamin Mahieu[a,*], Michel Visalli[a], Arnaud Thomas[b], Pascal Schlich[a]

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

ARTICLE INFO

Keywords:
Free-Comment Attack-Evolution-Finish (FC-AEF)
Open-ended questions
Temporal sensory method
Home Used Test (HUT)
Consumer study

ABSTRACT

Temporal Dominance of Sensations (TDS) and Temporal-Check-All-That-Apply (TCATA) are the most popular methods used with consumers for the temporal sensory characterization of a set of products. However, TDS and TCATA share the same limitation: they rely on a predefined and necessarily short list of descriptors. Free-Comment (FC) enables the sensory characterization of a set of products freed of any issue induced by the use of a list of descriptors, but for practical reasons collecting FC descriptions concurrently to the product intake is nearly impossible. Attack-Evolution-Finish (AEF) is an alternative to TDS and TCATA that replace concurrent by retrospective data collection. In AEF, subjects are asked to choose in a list one descriptor for each of the so-called periods: Attack, Evolution, and Finish. The paper introduced Free-Comment Attack-Evolution-Finish (FC-AEF) to extend FC to temporal sensory analysis where descriptor selections of AEF are replaced by FC descriptions. FC-AEF has been used at home with 63 consumers having tasted five dark chocolates. The data were analysed product-wise and period-wise and showed that FC-AEF enabled to provide temporal discrimination and characterization of the products. The product-wise analyses identified in each period the descriptors of each product enabling this discrimination. The period-wise analyses identified for each product the descriptors generating a temporal kinetic of its perception.

1. Introduction

Since it has been advocated that sensory perception is not a static phenomenon but rather a dynamic one (Lee & Pangborn, 1986), several methods have been developed to study the kinetic of sensations during the perception of a product. It is possible to distinguish two subcategories of temporal sensory methods: quantitative-based ones and qualitative-based ones. Among quantitative-based methods, we can mention Time-Intensity (Lee & Pangborn, 1986), Dual-Attribute Time-Intensity (Duizer, Bloom, & Findlay, 1996), Multi-Attribute Time-Intensity (Kuesten, Bi, & Feng, 2013), Progressive Profile (Jack, Piggott, & Paterson, 1994) and Sequential Profile (Methven et al., 2010). Quantitative-based methods require a trained panel, which implies a time-consuming and possibly expensive training period before starting product evaluations. Among qualitative-based temporal sensory methods, the two most popular are Temporal Dominance of Sensations (TDS) (Pineau, Cordelle, Imbert, Rogeaux, & Schlich, 2003; Pineau et al., 2009) and Temporal-Check-All-That-Apply (TCATA) (Castura, Antúnez, Giménez, & Ares, 2016). Contrary to quantitative-based methods, TDS and TCATA can be used with consumers without specific training (Jaeger et al., 2018; Rodrigues et al., 2016; Schlich, 2017).

During a TDS task, the subjects are asked to select among a predefined list of descriptors, which one is "dominant" at each time within a product intake (Pineau et al., 2003, 2009). A descriptor is considered as dominant from its selection until another descriptor is selected as being dominant instead. TCATA adopts another rational than TDS by enabling the subjects to select several descriptors at each time within a product intake (Castura et al., 2016). In practice, subjects select a descriptor when they judge it applicable and unselect a descriptor when they judge it no longer applicable. Both TDS and TCATA share the same limitation: they rely on a predefined and necessarily short list of descriptors (Jaeger et al., 2018; Pineau et al., 2012).

Establishing a list of descriptors is very tedious and represents a critical step for the relevance of the collected data as it may affect the results of the study (Ares et al., 2013; Pineau et al., 2012; Varela et al., 2018). Furthermore, several sources of bias induced by the use of a predefined list of descriptors have been reported in the literature. The list influences the subjects by suggesting descriptors that they would

* Corresponding author at: Centre des Sciences du Goût et de l'Alimentation, AgroSup Dijon, CNRS, INRAE, Université Bourgogne Franche-Comté, F-21000 Dijon, France.
E-mail address: benjamin.mahieu@inrae.fr (B. Mahieu).

https://doi.org/10.1016/j.foodqual.2020.104008
Received 17 April 2020; Received in revised form 19 June 2020; Accepted 23 June 2020
Available online 26 June 2020
0950-3293/

not think about otherwise (Coulon-Leroy, Symoneaux, Lawrence, Mehinagic, & Maitre, 2017; Kim, Hopkinson, van Hout, & Lee, 2017; Krosnick, 1999). Since the list contains only a limited number of descriptors, subjects may select descriptors that are close to what they perceive but not representing exactly what they actually perceive (Krosnick, 1999) and the collected data can be biased by the dumping effect (Varela et al., 2018). The first descriptors of the list (in the sense of presentation order) have a greater chance of being selected (Castura, 2009; Kim et al., 2017; Krosnick, 1999; Pineau et al., 2012).

Free-Comment (FC) (ten Kleij & Musters, 2003), as a response to open-ended questions, has proven itself an efficient method in characterizing and discriminating sets of products both with consumers and with experts (Lahne, Trubek, & Pelchat, 2014; Lawrence et al., 2013; ten Kleij & Musters, 2003) even out of the lab (Mahieu, Visalli, Thomas, & Schlich, 2020). As FC does not require a predefined list of descriptors, all the issues mentioned above do not longer hold. However, the FC method does not enable temporal sensory characterization.

For the products that have a relatively short tasting duration (say up to 45 s), collecting FC temporal descriptions in continuous time concurrently to the product intake as in TDS and TCATA is nearly impossible for practical reasons. Indeed, subjects should have first to identify the sensations they perceive within a complex signal, then think about the words that best describe these sensations and then finally transcript these words (handwriting, keyboard input, or voice recording) while staying focused on their perception. It would therefore not be reasonable to consider the data as being collected concurrently to the perception.

The recently introduced Attack-Evolution-Finish (AEF) method (Visalli, Mahieu, Thomas, & Schlich, 2020) proposes an alternative to continuous concurrent data collection. During an AEF task, subjects are asked to select retrospectively among a predefined list of descriptors which one they perceived during the so-called periods: Attack, Evolution, and Finish. The results obtained from AEF and TDS were compared in a study involving 120 consumers having evaluated five dark chocolates. AEF and TDS provided equivalent product discrimination and a very similar product characterization (Visalli et al., 2020).

The paper introduces the Free-Comment Attack-Evolution-Finish (FC-AEF), a method that integrates AEF and FC. In FC-AEF, the descriptor selection for each of the three periods (Attack, Evolution, and Finish) is replaced by an FC description, enabling a temporal sensory characterization without the issues induced by the use of a predefined list of descriptors.

The present study investigated whether consumers can successfully conduct an FC-AEF protocol at home and whether it enables the temporal characterization and discrimination of a set of products.

2. Material and methods

2.1. Participants

To create a situation as close as possible to an everyday consumption situation, the study took place at home with 63 naïve subjects (25 men and 38 women), 18 to 60 years old. The subjects were recruited from a population registered in the ChemoSens Platform's PanelSens database. This database has been declared to the relevant authority (Commission Nationale Informatique et Libertés—CNIL—n° d'autorisation 1148039). The subjects were consumers of dark chocolates at least once every two weeks and were rewarded for their participation in the study.

2.2. Products

Five dark chocolates provided by Barry Callebaut® were used for this study. They differed on their percentage of cocoa as well as on the origin of the cocoa used in the recipe. SDC has 54.5% of cocoa obtained from a mix of cocoa beans. BRA has 66.8% of cocoa coming from Brazil. EQU has 70.4% of cocoa coming from Ecuador. MAD has 67.4% of cocoa coming from Madagascar. SAO has 70% of cocoa coming from Sao Tomé. The chocolates were delivered to the subjects in sealed plastic containers in the form of callets (pucks of chocolates formulated for melting rather than baking). The subjects were invited to store the chocolates in a relatively cold place so that they did not melt or alter.

2.3. Data acquisition

2.3.1. General procedure

The subjects participated in five home-based sessions on their computers running TimeSens© software 2.0 (INRAE, Dijon, France). To access the sessions, the subjects simply had to click on a link sent to them by e-mail. In each session, consumers had to evaluate and describe only one product; it lasted approximately 5 min. The presentation of the products (and thus the sessions) was arranged following a William Latin square design. The minimum interval between two sessions was forced to be at least 24 h.

2.3.2. FC-AEF task

The instructions were given to the subjects at the beginning of the first session: “You are going to taste five chocolates. Each tasting will be separated from the previous one by at least 24 h. For each chocolate, you will be asked to describe the sensations you perceived during the tasting in the chronological order that you perceived these sensations. You will provide the descriptions using your own words.” An example was given to the subjects right after the instructions: “Example: At first, I perceived this chocolate sour and soft, then after a few moments I perceived it sour, sticky and woody, and at the end of the tasting I perceived it astringent, melting and sweet”. This example had the objective to inform the subjects that the same word could be used for several periods and that several different words could be used in the same period. This was underlined by the following sentence right after the example: “You can use the same words for several periods and several different words can be used in the same period”. This was underlined by the following sentence right after the example: “You can use the same words for several periods and several different words can be used in the same period”.

Fig. 1 shows the FC-AEF data collection screen. For each product evaluation, the following instruction was given to the subjects: “What sensations did you perceive during the tasting (textures, flavors, aromas, etc.) in chronological order? (Use your own words to answer)”. Three text areas corresponding to each period (Attack, Evolution, and Finish) were displayed on the screen. The text areas were organized on the screen so that the subjects filled the following sentence when describing their perception: “At first, I perceived this chocolate…, then after a few moments I perceived it…, and at the end of the tasting I

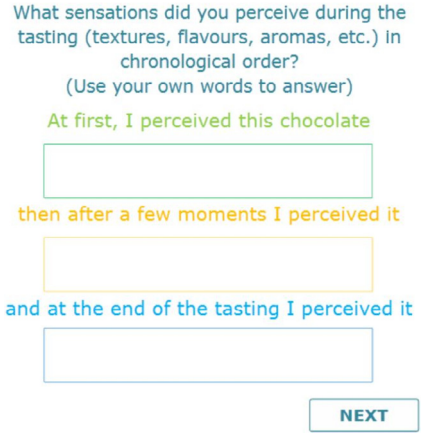


Fig. 1. FC-AEF data collection screen (translated from French).

perceived it…" (Visalli et al., 2020).

No particular restriction was given to the subjects on the manner of stating their descriptions. The subjects were forced to give at least one word within each period.

2.4. FC-AEF data treatment

As descriptions were collected in French, all the pre-treatments were performed in French. The analysed words resulting from the treatments have been translated into English for the present paper. The English-French correspondence of the analysed words can be found in the appendix.

All the FC-AEF data treatments were performed using R 3.5.1 (R Core Team, 2018). The lexicon provided with IRaMuTeQ© software (Ratinaud, 2014) was used for lemmatization and part-of-speech tagging. The data of the three periods were merged before applying the following pre-treatments. This merging was done only for the pre-treatments of the descriptions and to ensure that the data from each of the three periods were treated the same manner. The procedure used was the same one as described in Mahieu, Visalli, Thomas, and Schlich (2020) and summarized thereafter.

The descriptions were first cleaned, lemmatized, and filtered. Then, the words with similar meanings were grouped into latent-words relying on the chi-square-distance-based ascendant hierarchical classification.

Among all the words and latent words (simply called words hereafter for simplification), only those mentioned by at least 5% of the panel for at least one same product within at least one same period were retained for further analysis.

Finally, the number of times each remaining word was cited within each period for each product was computed at the panel level. Three contingency tables, one per period, containing the citation counts of each word for each product were built. These contingency tables will be referred subsequently as "product by word contingency tables". Five contingency tables, one per product, containing the citation counts of each word for each period were built. These contingency tables will be referred subsequently as "period by word contingency tables".

2.5. Data analyses

All analyses were performed using R 3.5.1 (R Core Team, 2018).

2.5.1. Panel behavior

The distributions of the number of analysed words (after pre-treatments) cited by each subject, for each product and each period as well as for the three periods aggregated were computed. For a given evaluation (product × subject), the number of analysed words for the three periods aggregated corresponds to the sum of citations of analysed words of the three periods. Thus, for the aggregated data, the same word can be cited more than once per evaluation. The mean, the mode, and the standard deviation of these four distributions were computed.

2.5.2. Contingency tables

The eight contingency tables (a "product by word contingency table" for each of the 3 periods [A, E and, F] and a "period by word contingency table" for each of the 5 products [SDC, BRA, EQU, MAD and, SAO]) were analysed the same manner following the procedure presented in Mahieu, Visalli, and Schlich (2020) and summarized thereafter. A chi-square test using a Monte Carlo approach (1000 simulations, $\alpha = 5\%$) was performed to investigate the significance of the dependence between products or periods and words. If the chi-square test was significant, a correspondence analysis (CA) was applied to the contingency table. The standard CA biplot was used to display the CA results. The number of significant CA axes was determined using the Monte-Carlo tests of dependence (1000 simulations, $\alpha = 5\%$). The confidence ellipses for the products or the periods coordinates in the CA space were computed with a total bootstrap procedure (1000 bootstrap samples, $\alpha = 5\%$) in which Procrustes rotations were performed on the significant axes. To assess relations between products or periods and words, Fisher's exact tests ($\alpha = 5\%$) per cell with a one-sided greater alternative hypothesis were conducted on the derived contingency table corresponding to significant axes. This contingency table is computed by reversing the CA computations on the significant axes (Mahieu, Visalli, & Schlich, 2020). To assess products or periods discrimination, a total bootstrap test ($\alpha = 5\%$) (Mahieu, Visalli, Thomas, & Schlich, 2020) was performed for each pair of products or periods on the significant axes.

3. Results

3.1. Panel behavior

Fig. 2 shows that the three periods had very similar distributions in terms of effective words cited. The number of effective words cited ranged from 0 to 4 (Attack period) or 5 (Evolution and Finish period). The mode of the three distributions was equal to 1, the mean was around 1.43 and the standard deviation ranged from 0.82 (Attack period) to 0.97 (Finish period). The standard deviation slightly increased from the Attack period to the Finish period.

For all periods aggregated, Fig. 2 (d) shows that the number of effective words cited for each subject and each product ranged from 0 to 10 with a mode of 4, a mean of 4.3, and a standard deviation of 1.96.

3.2. Product by word contingency tables

Table 1 shows that FC-AEF presented three significant axes for the Attack and the Evolution periods and only one significant axis for the Finish period. Therefore, a product by word significant dependence was detected in each period, though less complex in the Finish period.

Fig. 3 shows that the first dimension of the product configuration was very similar across the three periods and mostly opposed SDC to BRA with SAO, MAD, and EQU being placed between them. This first dimension seemed to be a gradient of strength induced by the opposition of strong and slight flavors. Fig. 3 (b) shows that the second dimension of the Attack period mostly opposed MAD to the other products. This dimension seemed to be a texture gradient of hardness. Fig. 3 (b) shows that the third dimension of the Attack period mostly opposed EQU and SAO. This dimension seemed to be a gradient of sweetness associated with a second gradient of hardness. Fig. 3 (d) shows that the second dimension of the Evolution period had high similarity with the third dimension of the Attack period, mostly opposing EQU and SAO. This dimension seemed to be a gradient of sweetness but it also showed an opposition between several flavors and textures. The third dimension of the Evolution period did not show an obvious interpretation.

The product discrimination was weaker at the Finish period as compared to the Attack and Evolution periods. The five products were discriminated for the Attack and Evolution periods but not for the Finish period, where only seven pairs of products out of ten were discriminated. Fig. 3 (e) suggests that the subjects only found large differences between SDC and the other products at the finish of the product perception. These latter seem not to have any particular characteristics distinguishing them from each other at the end of the intake.

Fig. 4 shows that the product discrimination into each period was driven by descriptors specific to the period. Indeed, the five products showed a kinetic of the characteristics that discriminate them from each other throughout the periods. From the Attack to the Evolution period, SDC lost its association with *crunchy_hard* and became associated with *fat*. From the Evolution to the Finish period, SDC lost its association with *fat* and became associated with *not_bitter* and *gentle_slight*. From the Attack to the Evolution period, BRA became associated with *spicy*. From

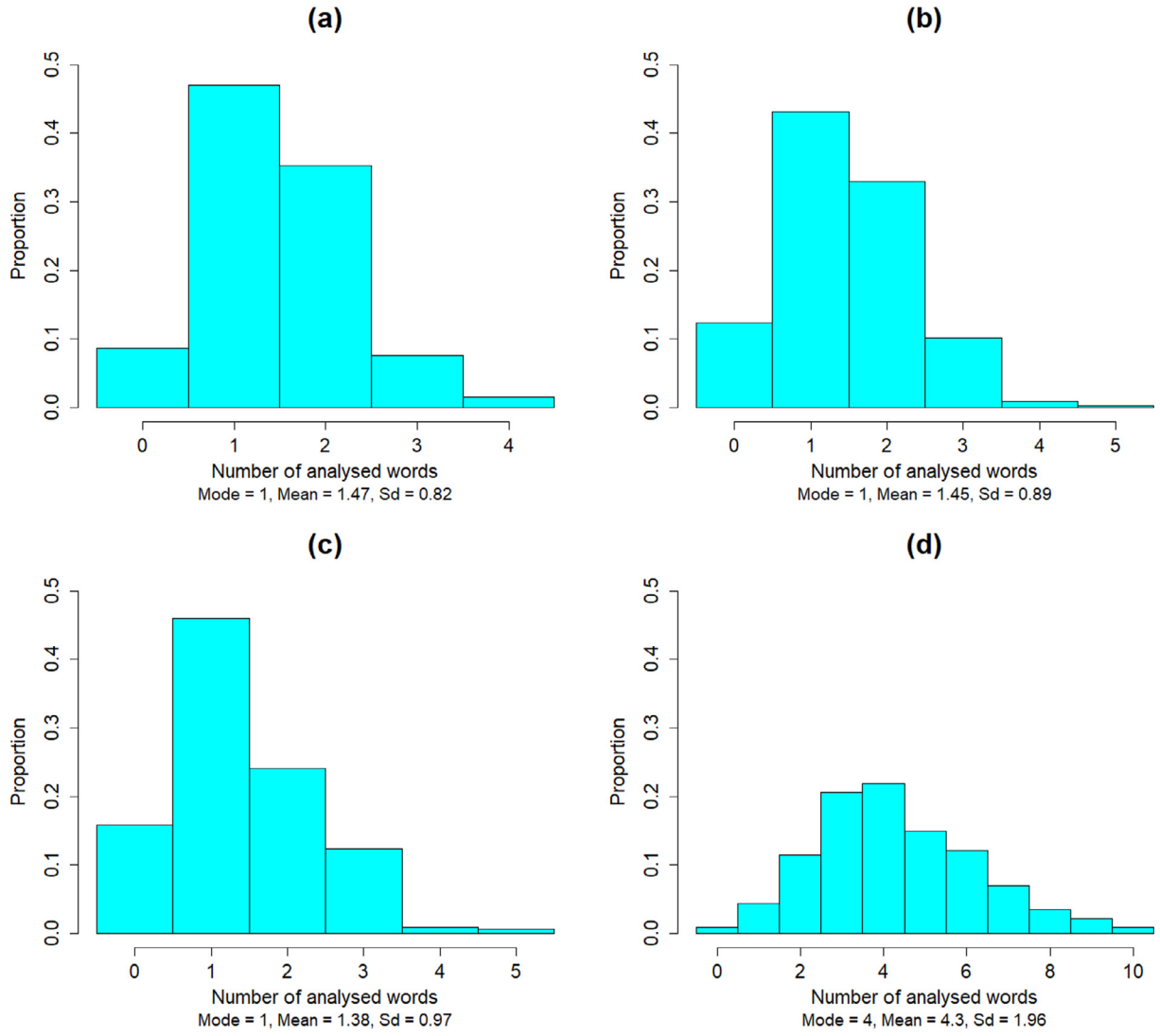


Fig. 2. Distributions of the number of analysed words (after pre-treatments) cited by each subject for each product for: (a) the Attack period, (b) the Evolution period, (c) the Finish period and (d) the three periods aggregated.

Table 1
p-Values of the test of dependence for each axis of each period.

Period	P-value: chi-square/ axis 1	P-value: axis 2	P-value: axis 3	P-value: axis 4
Attack	< 0.001	0.0019	0.0029	0.2257
Evolution	< 0.001	0.0119	0.0169	0.4725
Finish	< 0.001	0.1288	0.6443	0.6023

the Evolution to the Finish period, BRA lost its associations with *spicy*, *strong_intense_powerful,* and *bitter*. At the Finish period, no significant association was found between BRA and the descriptive words. From the Attack to the Evolution period, EQU lost its associations with *not_sweet*. At the Evolution and Finish periods, no significant association was found between EQU and the descriptive words. From the Attack to the Evolution period, MAD lost its associations with *melting_smooth_creamy* and *soft*. At the Evolution and Finish periods, no significant association was found between MAD and the descriptive words. From the Attack period to the Evolution period, SAO became associated with *bitter*. At the Attack and Finish periods, no significant association was found between SAO and the descriptive words. The results concerning the Finish period shown by Fig. 4 tends to confirm that the subjects did not find large differences between the products at the Finish period except for SDC that was associated with four words. Indeed, the *sweet* and *gentle_slight* characteristics of SDC seem to increase over time as compared to the other products.

3.3. Period by word contingency tables

For the five products, the two axes of the CA performed on their respective period by word contingency table were highly significant. The largest of these p-values was 0.0029. This shows that for each product, the three periods were discriminated from each other.

Fig. 5 shows results in line with the tests of dependence: all periods were discriminated from each other for all products. For each of them, the period configurations were similar: the first axis mostly opposed the Attack period to the Finish period while the second axis opposed the Evolution period to the Attack and Finish periods. Words related to the texture (e.g. *crunchy_hard*) and words related to the end of perception (e.g. *long_tasting*) seemed to be the most important drivers of the period configuration for all the products. However, these main drivers were associated with flavors and aromas descriptions that depended on the period for each product.

Fig. 6 confirms that the period discrimination was mainly due to the texture and the end of perception descriptions. Indeed, *crunchy_hard* was associated with the Attack period for all the products,

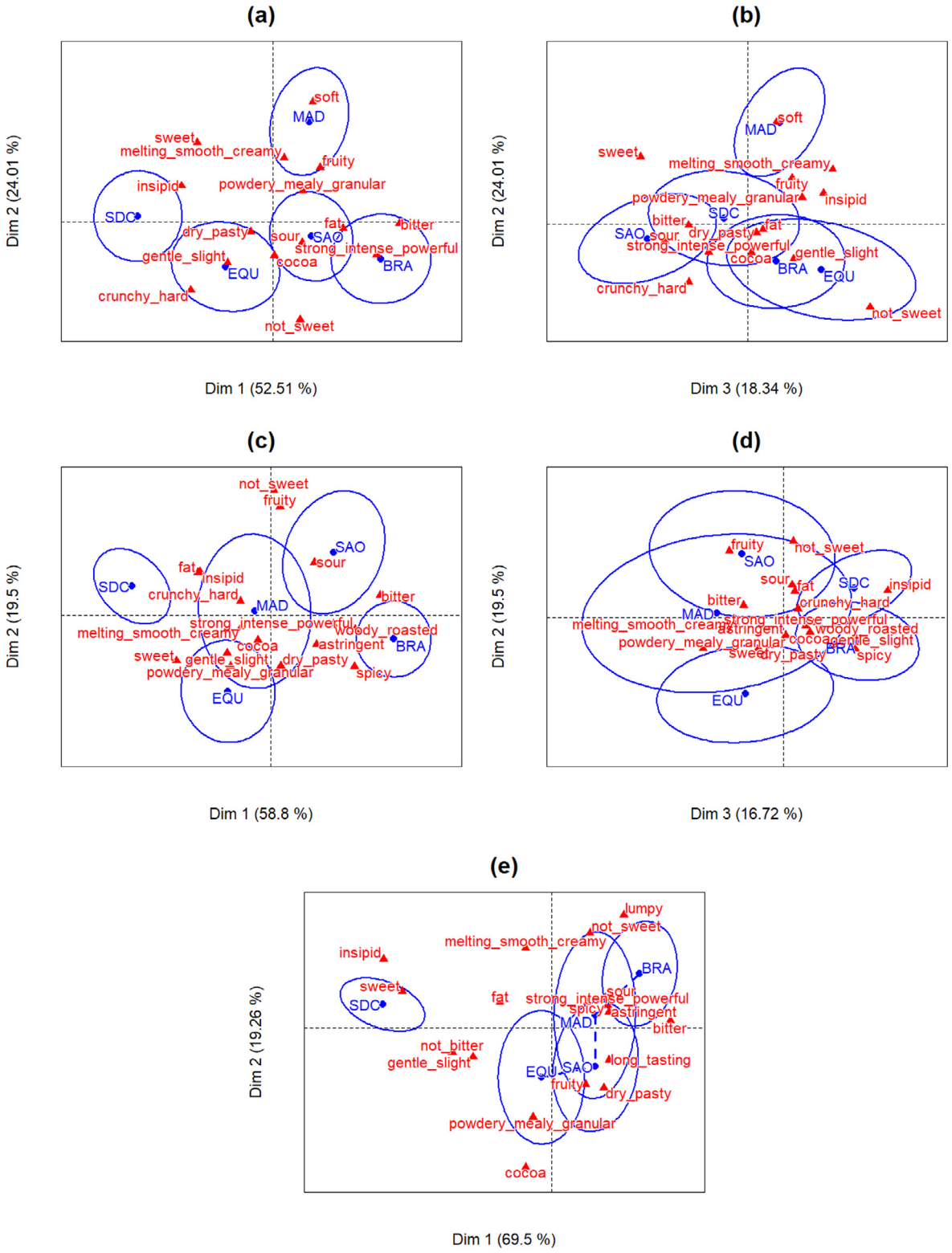


Fig. 3. Correspondence analysis standard biplot of product by word contingency tables by period: (a) Attack axes 1–2, (b) Attack axes 3–2, (c) Evolution axes 1–2, (d) Evolution axes 3–2 and (e) Finish axes 1–2. Two products linked by a dashed line are not significantly different (total bootstrap test, $\alpha = 5\%$).

melting_smooth_creamy was associated with the Evolution period for all the products except BRA, and *long_tasting* was associated with the Finish period of all the products except SDC. This kinetic was common to all the products.

Fig. 6 suggests that all products showed a temporal kinetic since the periods had different characteristics relatively to each other. SDC showed a texture kinetic, being perceived more often *crunchy_hard* and dry_pasty at the Attack period and then *fat* and *melting_smooth_creamy* at the Evolution period. SDC was specifically more described as *not_bitter* at the Finish period. BRA showed a multi-modal kinetic, being perceived more often *crunchy_hard* and *powdery_mealy_granular* at the Attack period, then *woody_roasted* at the Evolution period and finally

	(a)					(b)					(c)				
	SDC	BRA	EQU	MAD	SAO	SDC	BRA	EQU	MAD	SAO	SDC	BRA	EQU	MAD	SAO
astringent	0	0	0	0	0	0	4.8	1.6	3.2	1.6	0	3.2	1.6	6.3	3.2
bitter	4.8	28.6	7.9	17.5	25.4	7.9	38.1	20.6	27	38.1	9.5	38.1	30.2	39.7	31.7
cocoa	11.1	12.7	9.5	6.3	9.5	12.7	12.7	11.1	11.1	7.9	9.5	1.6	12.7	7.9	15.9
crunchy_hard	38.1	17.5	33.3	12.7	33.3	4.8	1.6	3.2	1.6	3.2	0	0	0	0	0
dry_pasty	14.3	11.1	12.7	9.5	11.1	3.2	6.3	7.9	6.3	4.8	4.8	11.1	12.7	11.1	15.9
fat	0	6.3	1.6	3.2	3.2	14.3	3.2	4.8	11.1	4.8	7.9	4.8	4.8	1.6	3.2
fruity	0	3.2	1.6	4.8	1.6	3.2	1.6	0	6.3	7.9	1.6	3.2	3.2	3.2	7.9
gentle_slight	22.2	15.9	25.4	12.7	15.9	22.2	15.9	19	12.7	11.1	23.8	7.9	19	17.5	12.7
insipid	17.5	4.8	11.1	9.5	1.6	11.1	3.2	1.6	1.6	1.6	12.7	0	1.6	1.6	0
long_tasting	0	0	0	0	0	0	0	0	0	0	1.6	7.9	6.3	4.8	9.5
lumpy	0	0	0	0	0	0	0	0	0	0	0	6.3	0	1.6	1.6
melting_smooth_creamy	7.9	9.5	15.9	19	9.5	19	14.3	20.6	25.4	17.5	9.5	7.9	3.2	6.3	4.8
not_bitter	0	0	0	0	0	0	0	0	0	0	6.3	0	3.2	0	1.6
not_sweet	1.6	7.9	6.3	0	0	6.3	3.2	0	3.2	9.5	1.6	4.8	0	4.8	1.6
powdery_mealy_granular	1.6	4.8	4.8	6.3	3.2	3.2	0	7.9	6.3	3.2	1.6	0	4.8	1.6	1.6
soft	1.6	3.2	0	7.9	1.6	0	0	0	0	0	0	0	0	0	0
sour	1.6	3.2	1.6	1.6	6.3	1.6	4.8	0	3.2	4.8	1.6	6.3	1.6	4.8	6.3
spicy	0	0	0	0	0	0	9.5	1.6	1.6	1.6	3.2	6.3	4.8	9.5	4.8
strong_intense_powerful	0	14.3	4.8	6.3	12.7	3.2	19	7.9	7.9	12.7	4.8	12.7	9.5	14.3	9.5
sweet	28.6	6.3	7.9	17.5	19	28.6	9.5	27	20.6	12.7	34.9	7.9	14.3	14.3	11.1
woody_roasted	0	0	0	0	0	0	4.8	1.6	0	3.2	0	0	0	0	0

Fig. 4. Words by product percentages of citation across the panel for the period: (a) Attack, (b) Evolution and (c) Finish. Cells highlighted in green show the results of Fisher's exact tests ($\alpha = 5\%$). Grey cells correspond to words cited in another period than the one considered. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

lumpy and *long_tasting* at the Finish period. EQU showed the strongest kinetic and a very interesting one. It was perceived more often *crunchy_hard*, *insipid,* and *not_sweet* at the Attack period, then *sweet* and *melting_smooth_creamy* at the Evolution period and finally, *bitter* and *long_tasting* at the Finish period. MAD also presented an interesting kinetic. It was perceived more often *crunchy_hard*, *insipid* and *soft* at the Evolution period, then *fat* and *melting_smooth_creamy* at the Evolution period and finally, *bitter*, *long_tasting* and *spicy* at the Finish period. SAO only showed a slight kinetic, being perceived more often *crunchy_hard* at the Attack period, then *melting_smooth_creamy* and *not_sweet* at the Evolution period, and finally, *long_tasting* at the Finish period.

4. Discussion

The temporal aspect of the FC-AEF task seems to have been understood by the subjects. Indeed, the words related to texture aspects (e.g. *crunchy_hard*) were only mentioned in the Attack period, some sensations related to the end of the perception (e.g. long_tasting) were only mentioned in the Finish period.

The empirical results of Fig. 1 show that on average only one word and half are kept as an analysed word by period for each evaluation (subject × product). This results in an average of 4.3 analysed words per evaluation (all periods aggregated), which is not a huge increase as compared to the three words per evaluation imposed in the AEF method. However, this might be depending on the product type. It is also interesting to note that for the three periods, about 10% of the evaluations were associated with zero analysed words. This does not mean that subject did not report descriptors, but that the pre-treatment removes these descriptors. Indeed, some descriptions were composed of only hedonic words (e.g. "good taste"), some others were composed of low cited words (e.g. "salty") and the others were composed of uninformative words (e.g. "aromas").

The results of the analyses of product by word contingency tables enabled to identify the periods of the product intake that enabled the products to be discriminated as well as the characteristics of each product leading to this discrimination. The first dimension remaining stable across all periods suggests that the main latent dimension of discrimination is independent of time for this set of products. This dimension was a gradient of strength of the chocolates and did not evolve across periods of the product intake.

The results of the CA applied on the period by word contingency tables presented a particular period configuration for all the products. The first axis systematically opposed the Attack period to the Finish period and the second axis systematically opposed the Evolution period to the Attack and Finish periods. It is mainly due to the texture and end of perception descriptions of the products. Indeed, it seems that almost all products were perceived *crunchy_hard* at the beginning, *melting_smooth_creamy* during the consumption and *long_tasting* at the end of the perception, at least for several subjects. This particular period configuration is likely to occur for all types of products that present an obvious kinetic of some sensations throughout the intake (e.g. textures).

Concerning the analyses of period by word contingency tables, the particular case of the product MAD is interesting: at the Attack period, two words with opposite meaning, namely *crunchy_hard* and *soft*, significantly characterized the product. It could be explained by the fact that from a subject to another, the range of time of the Attack and Evolution periods were not the same. It could also be that this product was first *crunchy_hard* and right after *soft*, leading some subjects to

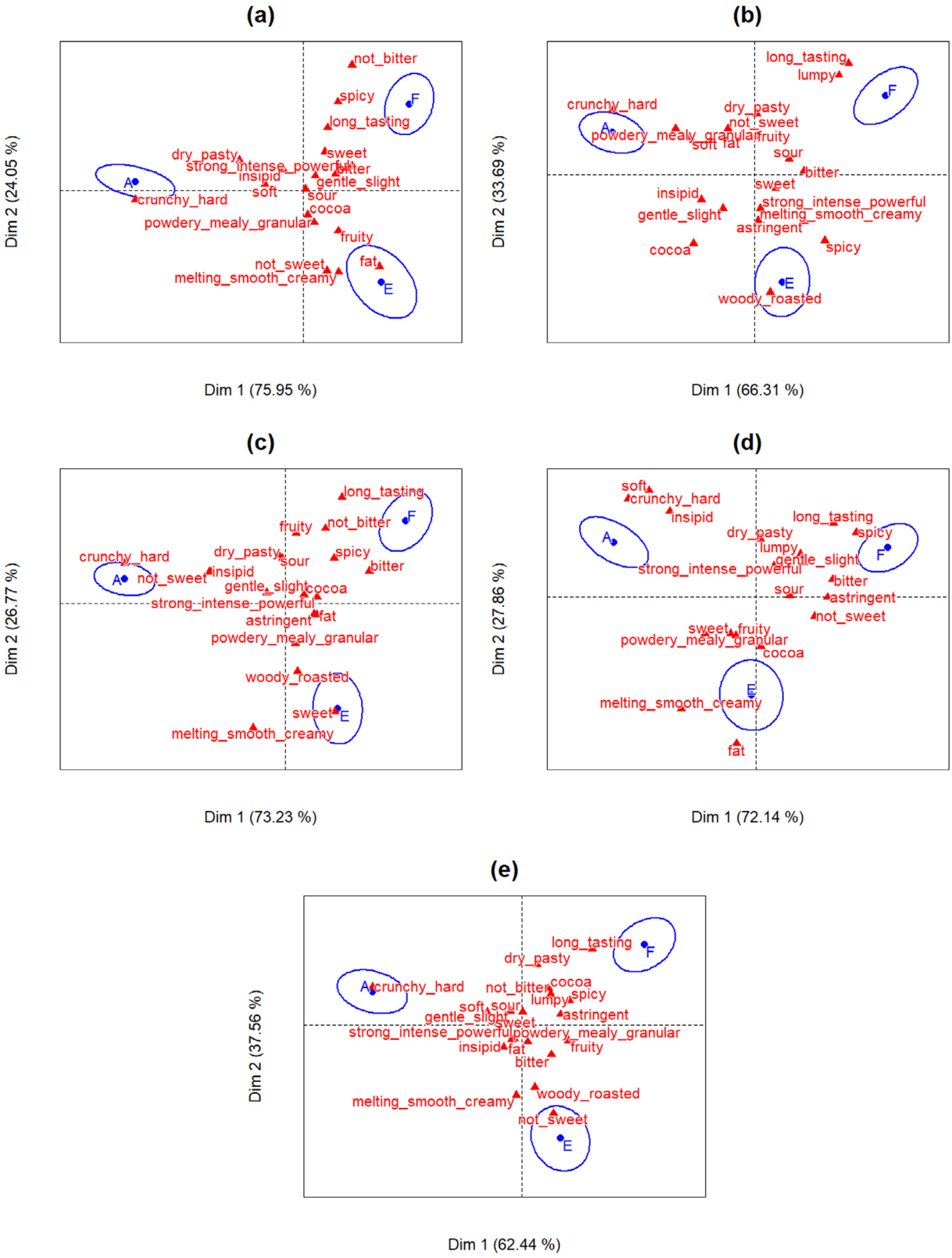


Fig. 5. Correspondence analysis standard biplot of period by word contingency tables of the product: (a) SDC, (b) BRA, (c) EQU, (d) MAD, (e) SAO.

describe it as *soft* and others as *crunchy_hard*. Another explanation would be that, depending on their references of black chocolate, some subjects perceived it *crunchy_hard* and some others *soft*. A mixture of these phenomena is likely to be what had happened. Anyhow, investigating individual representations of the three AEF periods would be of great interest, especially the range of time considered for each AEF period.

If a temporal sensory method relying on a predefined list of descriptors had been used instead of FC-AEF to characterize this set of products, a limited number of descriptors would have been used. As the

	(a)			(b)			(c)			(d)			(e)		
	A	E	F	A	E	F	A	E	F	A	E	F	A	E	F
astringent	0	0	0	0	4.8	3.2	0	1.6	1.6	0	3.2	6.3	0	1.6	3.2
bitter	4.8	7.9	9.5	29	38.7	38.7	7.9	20.6	30.2	17.5	27	39.7	25.4	38.1	31.7
cocoa	11.1	12.7	9.5	12.9	12.9	1.6	9.5	11.1	12.7	6.3	11.1	7.9	9.5	7.9	15.9
crunchy_hard	38.1	4.8	0	17.7	1.6	0	33.3	3.2	0	12.7	1.6	0	33.3	3.2	0
dry_pasty	14.3	3.2	4.8	11.3	6.5	11.3	12.7	7.9	12.7	9.5	6.3	11.1	11.1	4.8	15.9
fat	0	14.3	7.9	6.5	3.2	4.8	1.6	4.8	4.8	3.2	11.1	1.6	3.2	4.8	3.2
fruity	0	3.2	1.6	3.2	1.6	3.2	1.6	0	3.2	4.8	6.3	3.2	1.6	7.9	7.9
gentle_slight	22.2	22.2	23.8	16.1	16.1	8.1	25.4	19	19	12.7	12.7	17.5	15.9	11.1	12.7
insipid	17.5	11.1	12.7	4.8	3.2	0	11.1	1.6	1.6	9.5	1.6	1.6	1.6	1.6	0
long_tasting	0	0	1.6	0	0	8.1	0	0	6.3	0	0	4.8	0	0	9.5
lumpy	0	0	0	0	0	6.5	0	0	0	0	0	1.6	0	0	1.6
melting_smooth_creamy	7.9	19	9.5	9.7	14.5	8.1	15.9	20.6	3.2	19	25.4	6.3	9.5	17.5	4.8
not_bitter	0	0	6.3	0	0	0	0	0	3.2	0	0	0	0	0	1.6
not_sweet	1.6	6.3	1.6	8.1	3.2	4.8	6.3	0	0	0	3.2	4.8	0	9.5	1.6
powdery_mealy_granular	1.6	3.2	1.6	4.8	0	0	4.8	7.9	4.8	6.3	6.3	1.6	3.2	3.2	1.6
soft	1.6	0	0	3.2	0	0	0	0	0	7.9	0	0	1.6	0	0
sour	1.6	1.6	1.6	3.2	4.8	6.5	1.6	0	1.6	1.6	3.2	4.8	6.3	4.8	6.3
spicy	0	0	3.2	0	9.7	6.5	0	1.6	4.8	0	1.6	9.5	0	1.6	4.8
strong_intense_powerful	0	3.2	4.8	14.5	19.4	12.9	4.8	7.9	9.5	6.3	7.9	14.3	12.7	12.7	9.5
sweet	28.6	28.6	34.9	6.5	9.7	8.1	7.9	27	14.3	17.5	20.6	14.3	19	12.7	11.1
woody_roasted	0	0	0	0	4.8	0	0	1.6	0	0	0	0	0	3.2	0

Fig. 6. Words by period percentages of citation across the panel for the product: (a) SDC, (b) BRA, (c) EQU, (d) MAD, (e) SAO. Cells highlighted in green show the results of Fisher's exact tests ($\alpha = 5\%$). Grey cells correspond to words cited for another product than the one considered. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

product space was the same as in Visalli et al. (2020), the list would likely have also been the same, or at least very close. This list contains the following descriptors: *Dry, Floral, Sweet, Bitter, Fat, Melting, Sour, Astringent, Woody, Sticky, Cocoa,* and *Fruity*. Except for the descriptors *Floral* and *Sticky*, all the descriptors contained in this list were used by the subjects in their descriptions. This means that subjects were able to generate an appropriate list of words to be used for describing this set of products. However, it is interesting to note that *astringent* and *cocoa* were only sparsely employed relatively to when they are proposed in a list (Visalli et al., 2020). *Astringent* maybe not a well-known word by the consumers and *cocoa* might sounds too obvious for several subjects when they do not belong to a list. Compared to the pre-defined list, subjects also provided nine additional words that seem very important for the description of this set of products: *crunchy_hard, insipid, strong_intense_powerful, soft, spicy, gentle_slight, powdery_mealy_granular, long_tasting* and *lumpy*. This additional information suggests that using a predefined list would have resulted in a loss of information. It was expected that the descriptor "crunchy_hard" appeared in the descriptions since "Crunchy" was originally part of the list used in Visalli et al. (2020). However, several TDS studies exhibited a systematic selection of this descriptor at the beginning of the perception for every black chocolate, thus limiting the selection of other descriptors at this stage of the perception. For this reason, it was removed from the list of descriptors. Since AEF limits the description of the Attack period to a single descriptor, it was even more crucial not to include "Crunchy" in the list used in Visalli et al. (2020) to avoid obtaining trivial descriptions of the Attack period. However, because FC-AEF does not share this limit on the number of descriptors with AEF, it was able to highlight "crunchy_hard" as a key descriptive word of first chewing cycles that discriminated between products and periods, which is a nice addition compared to AEF.

The variability of the number of terms that can be selected within each period makes FC-AEF closer to TCATA than TDS or AEF, which both forces the subjects to select one descriptor at a given time or period. However, by being retrospective, FC-AEF, as well as AEF, are different from TDS and TCATA, which are concurrent time-dependent measures. As discussed in Visalli et al. (2020), AEF, and thus FC-AEF too, rely on short-term memory while it is hoped that in TDS and TCATA subjects react more instinctively.

In this paper, two approaches to analyse the FC-AEF data have been proposed: product-wise and period-wise. In the product-wise approach, products are compared by period, while in the period-wise approach, periods are compared by product. These two approaches are complementary. For example, the product-wise approach informs that the product SDC was described sweeter than the other products in every period, while the period-wise approach informs that *sweet* was not used more often in a period than another for characterizing SDC. Depending on the problematic of the user, one of the approaches can be more appropriate than the other does. The product-wise approach is more appropriate if the study aims to investigate the differences between

products at specific steps of the product perception. The period-wise approach is more appropriate if it is assumed that the temporality of the perception may be different among products.

FC-AEF has been designed for temporal sensory characterization purposes. It is a suitable method when one wants to avoid the issues induced by the use of a predefined list of descriptors and when the temporal precision provided by list-based methods like TDS or TCATA is not crucial. Using FC-AEF implies losing a part the temporal precision provided by list-based methods but as a counterpart provides several benefits: descriptions are spontaneous, rich and precise, the dumping effect and the risk of missing key information are discarded and no limitations on the number of descriptors used in the descriptions exists. Further, from a practical point of view, FC-AEF also provides some benefits: no pre-tests for establishing a list of descriptors are required and the task does not need to be explained to the consumers since it is spontaneous. FC-AEF can also be considered as a relevant alternative to static FC to raise awareness of the subjects on the temporal kinetic of their perception in every application where static FC is suitable. The benefit of FC-AEF over static FC is that it enables to highlight the kinetics of the perception if any. If no kinetics exists, then FC-AEF data can be seen as static FC data and treated as such, since it can be expected that splitting the descriptions into three temporal periods does not flaw the overall description of the products.

5. Conclusion

This paper introduced a new temporal sensory method called Free-Comment Attack-Evolution-Finish (FC-AEF). This method is a combination of the Free-Comment and the Attack-Evolution-Finish methods in which for each of the so-called periods (Attack, Evolution, and Finish), subjects are asked to provide a Free-Comment description instead of selecting a descriptor in a predefined list. FC-AEF was used to collect temporal sensory perceptions of dark chocolates with consumers at home. The data collected were analysed product-wise and period-wise. The product-wise analysis identified in each period the descriptors characterizing each product, while the period-wise analysis identifies for each product the descriptors generating a temporal kinetic of its perception. FC-AEF provides sensory analysts with a new tool for investigating the temporal sensory perception of products by consumers with no need of establishing a predefined list of descriptors, which enables shunting this tedious part and removing all possible issues and biases due to the use of a predefined list.

CRediT authorship contribution statement

Benjamin Mahieu: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Michel Visalli:** Conceptualization, Methodology, Software, Validation, Resources, Writing - review & editing. **Arnaud Thomas:** Conceptualization, Validation, Writing - review & editing. **Pascal Schlich:** Conceptualization, Validation, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition.

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.

Appendix. English-French correspondence of the analysed words.

English	French
astringent	astringent
bitter	amer
cocoa	cacao
crunchy_hard	croquant_dur
dry_pasty	sec_pâteux
fat	gras
fruity	fruité
gentle_slight	doux_léger
insipid	fade
long_tasting	long_en_bouche
lumpy	âpre
melting_smooth_creamy	fondant_onctueux_crémeux
not_bitter	pas_amer
not_sweet	pas_sucré
powdery_mealy_granular	poudreux_farineux_granuleux
soft	mou
sour	acide
spicy	épicé
strong_intense_powerful	fort_intense_puissant
sweet	sucré
woody_roasted	boisé_torréfié

References

Ares, G., Jaeger, S. R., Bava, C. M., Chheang, S. L., Jin, D., Gimenez, A., et al. (2013). CATA questions for sensory product characterization: Raising awareness of biases. *Food Quality and Preference, 30*(2), 114–127.

Castura, J. C. (2009). Do panellists donkey vote in sensory choose-all-that-apply questions? In, 8th Pangborn Sensory Science Symposium, July 26-30. Florence, Italy.

Castura, J. C., Antúnez, L., Giménez, A., & Ares, G. (2016). Temporal check-all-that-apply (TCATA): A novel dynamic method for characterizing products. *Food Quality and Preference, 47*, 79–90.

Coulon-Leroy, C., Symoneaux, R., Lawrence, G., Mehinagic, E., & Maitre, I. (2017). Mixed Profiling: A new tool of sensory analysis in a professional context. Application to wines. *Food Quality and Preference, 57*, 8–16.

Duizer, L. M., Bloom, K., & Findlay, C. J. (1996). Dual-attribute time-intensity measurement of sweetness and peppermint perception of chewing gum. *Journal of Food Science, 61*(3), 636–638.

Jack, F. R., Piggott, J. R., & Paterson, A. (1994). Analysis of textural changes in hard cheese during mastication by progressive profiling. *Journal of Food Science, 59*(3), 539–543.

Jaeger, S. R., Alcaire, F., Hunter, D. C., Jin, D., Castura, J. C., & Ares, G. (2018). Number of terms to use in temporal check-all-that-apply studies (TCATA and TCATA Fading) for sensory product characterization by consumers. *Food Quality and Preference, 64*, 154–159.

Kim, I.-A., Hopkinson, A., van Hout, D., & Lee, H.-S. (2017). A novel two-step rating-based 'double-faced applicability' test. Part 1: Its performance in sample discrimination in comparison to simple one-step applicability rating. *Food Quality and Preference, 56*,

189–200.
Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537–567.
Kuesten, C., Bi, J., & Feng, Y. (2013). Exploring taffy product consumption experiences using a multi-attribute time–intensity (MATI) method. *Food Quality and Preference, 30*(2), 260–273.
Lahne, J., Trubek, A. B., & Pelchat, M. L. (2014). Consumer sensory perception of cheese depends on context: A study using comment analysis and linear mixed models. *Food Quality and Preference, 32*, 184–197.
Lawrence, G., Symoneaux, R., Maitre, I., Brossaud, F., Maestrojuan, M., & Mehinagic, E. (2013). Using the free comments method for sensory characterisation of Cabernet Franc wines: Comparison with classical profiling in a professional context. *Food Quality and Preference, 30*(2), 145–155.
Lee, W. E., III, & Pangborn, R. M. (1986). Time-intensity: The temporal aspects of sensory perception. *Food Technology, 40*(11), 71–78.
Mahieu, B., Visalli, M., & Schlich, P. (2020). 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. (2020). Free-comment outperformed check-all-that-apply in the sensory characterisation of wines with consumers at home. *Food Quality and Preference, 84*.
Methven, L., Rahelu, K., Economou, N., Kinneavy, L., Ladbrooke-Davis, L., Kennedy, O. B., et al. (2010). The effect of consumption volume on profile and liking of oral nutritional supplements of varied sweetness: Sequential profiling and boredom tests. *Food Quality and Preference, 21*(8), 948–955.
Pineau, N., Cordelle, S., Imbert, A., Rogeaux, M., & Schlich, P. (2003). Dominance temporelle des sensations – Codage et analyse d'un nouveau type de données sensorielles. In, 35èmes Journées de Statistiques, 2-6th June. Lyon, France.
Pineau, N., de Bouillé, A. G., Lepage, M., Lenfant, F., Schlich, P., Martin, N., et al. (2012). Temporal Dominance of Sensations: What is a good attribute list? *Food Quality and Preference, 26*(2), 159–165.
Pineau, N., Schlich, P., Cordelle, S., Mathonnière, C., Issanchou, S., Imbert, A., et al. (2009). Temporal Dominance of Sensations: Construction of the TDS curves and comparison with time–intensity. *Food Quality and Preference, 20*(6), 450–455.
R Core Team. (2018). R: A language and environment for statistical computing. In. Vienna, Austria: R Foundation for Statistical Computing.
Ratinaud, P. (2014). IRaMuTeQ : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. In. France.
Rodrigues, J. F., Souza, V. R.d., Lima, R. R., Carneiro, J. D. D. S., Nunes, C. A., & Pinheiro, A. C. M. (2016). Temporal dominance of sensations (TDS) panel behavior: A preliminary study with chocolate. *Food Quality and Preference, 54*, 51–57.
Schlich, P. (2017). Temporal Dominance of Sensations (TDS): A new deal for temporal sensory analysis. *Current Opinion in Food Science, 15*, 38–42.
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.
Varela, P., Antúnez, L., Carlehög, M., Alcaire, F., Castura, J. C., Berget, I., et al. (2018). What is dominance? An exploration of the concept in TDS tests with trained assessors and consumers. *Food Quality and Preference, 64*, 72–81.
Visalli, M., Mahieu, B., Thomas, A., & Schlich, P. (2020). Concurrent vs. retrospective temporal data collection: Attack-evolution-finish as a simplification of temporal dominance of sensations? *Food Quality and Preference* In Press.

C. Drivers of liking identification and ideal product characterization: Ideal-Free-Comment paired with liking scoring

Article in revision for Food Quality and Preference:

Identifying drivers of liking and characterizing the ideal product thanks to Free-Comment

Benjamin Mahieu*, Michel Visalli, Pascal Schlich

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

INRAE, PROBE research infrastructure, ChemoSens facility, F-21000 Dijon, France

Reference:

Mahieu, B., Visalli, M., & Schlich, P. (2021). Identifying drivers of liking and characterizing the ideal product thanks to Free-Comment. Manuscript in revision for Food Quality and Preference.

Title

Identifying drivers of liking and characterizing the ideal product thanks to Free-Comment

Authors

Benjamin Mahieu[1,2], Michel Visalli[1,2], Pascal Schlich[1,2]

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

[2]INRAE, PROBE research infrastructure, ChemoSens facility, F-21000 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.

INRAE, PROBE research infrastructure, ChemoSens facility, F-21000 Dijon, France

Highlights

- Drivers of liking can be identified from Free-Comment (FC) data
- Ideal-Free-Comment (IFC) enables to characterize the ideal product with FC
- IFC enables to locate the ideal product on the FC product map
- Drivers of liking based on FC data and IFC are complementary tools

Abstract

Consumers' hedonic appreciation is important for the commercial success of a product. To formulate appreciated products, sensory and hedonic data of some existing products are often linked to each other. Because existing products represent only a limited sensory space of investigation, asking consumers to characterize their ideal product can provide relevant additional information to understand their preferences. First, the paper investigates whether sensory drivers of liking can be derived from linking Free-Comment (FC) and hedonic data.

Second, Ideal-Free-Comment (IFC) is introduced. IFC instructs consumers to describe actual products and then their ideal product thanks to FC. IFC paired with liking scoring was used in a home-used test with 483 consumers each evaluating from 1 to 14 (5.71 on average) cooked hams from a list of 30 hams representative of the French market. Based on a mixed linear model, relevant drivers of liking were identified from FC data. The panel's average ideal product was consistent with the drivers of liking. Since descriptors with opposite meanings characterized individual ideal products, a consumer segmentation based on their ideal product was performed and resulted in two segments. The two segments' ideal products mainly differed regarding their flavor. Drivers of liking and the ideal product of the smaller segment (≈ 15% of the consumers) were not well consistent suggesting this was a noise segment. Drivers of liking based on FC data and IFC are complementary tools to understand consumers' hedonic appreciation without the use of a pre-established list of descriptors.

Keywords

- Open-ended questions
- Drivers of liking
- Ideal-Free-Comment (IFC)
- Consumer segmentation
- Cooked ham
- Home Used Test (HUT)

1. Introduction

Consumers' hedonic appreciation is one of the most important drivers of the commercial success of a product. It is most often investigated using hedonic tests in which a panel of consumers is instructed to score their overall liking of products. Since liking is a function of the products' sensory characteristics (Lagrange & Norback, 1987), investigating these characteristics is necessary to understand liking and formulate appreciated products. For this reason, hedonic tests are often performed conjointly to the sensory characterization of the products. Because consumers were claimed not to be able to provide valid nor reliable sensory characterization (Ares & Varela, 2017; Lawless & Heymann, 1999; Meilgaard, Civille, & Carr, 1991; Stone & Sidel, 1993), this sensory characterization used to be performed by sensory profiling using a trained panel.

Several methodologies have been developed to link sensory and hedonic data among which preference mapping techniques (Carroll, 1972; Danzart, 2009; Greenhoff & MacFie, 1994; McEwan, 1996; Schlich & McEwan, 1992) are likely the most popular. Two major approaches can be distinguished among preference mapping techniques: internal preference mapping and external preference mapping. They mainly differ in the point of view they adopt (van Kleef, van Trijp, & Luning, 2006). Internal preference mapping puts the focus on the hedonic data: the product space is obtained from liking scores and the sensory descriptor scores are regressed into this space. On the contrary, external preference mapping puts the focus on the sensory data: the product space is obtained from sensory descriptor scores and the individual liking scores are regressed into this space. Worch (2013) proposed the so-called prefMFA method that uses Multiple Factor Analysis (Escofier & Pagès, 1994) to determine the shared dimensions between sensory and hedonic data.

During the last recent years, the affirmation upon which consumers are unable to provide valid or reliable sensory characterization has been reconsidered. One of the main reasons is that trained panels might consider descriptors and variations that are irrelevant to consumers (Ares & Varela, 2017; ten Kleij & Musters, 2003). In addition, several consumer methods were claimed to obtain more or less similar information as the one provided by sensory profiling in practical applications (Ares & Varela, 2017; Valentin, Chollet, Lelièvre, & Abdi, 2012; Varela & Ares, 2012). Among these consumer methods, some were specifically designed to understand preferences and to link hedonic data with consumer sensory data. Notably, Just-About-Right (JAR) scales (see for example Popper (2014)) and Check-All-That-Apply (CATA) (Adams, Williams, Lancaster, & Foley, 2007) paired with hedonic data collection and penalty-lift analysis (Meyners, Castura, & Carr, 2013) belong to these methods. Other methods sharing the same objective can be mentioned such as Preferred Attribute Elicitation (Grygorczyk, Lesschaeve, Corredig, & Duizer, 2013), preference mapping based on Sorting (Faye et al., 2006), and preference mapping based on CATA (Dooley, Lee, & Meullenet, 2010).

Previous methodologies intend to understand the sensory characteristics that drive the liking and the disliking of the products through the study of some existing products, which necessarily restricts the sensory space investigated. This limitation can affect the conclusions drawn since the ideal product does not necessarily lie within the product space (van Trijp, Punter, Mickartz, & Kruithof, 2007). Indeed, since only a limited number of products are presented to consumers then only a limited number of combinations of sensory characteristics are represented and evaluated. To circumvent this limitation, the Ideal-Profile-Method (IPM) (Moskowitz, 1972;

van Trijp et al., 2007; Worch, Lê, Punter, & Pagès, 2013) was proposed. In IPM, consumers are instructed to rate the products on several descriptors from a pre-established list using intensity scales. Right after the evaluation of every actual product, consumers are instructed to do the same task but considering a virtual ideal product. The idea is that the consumers provide for each descriptor the rating they would have found ideal in the previous actual product. Recently, characterizing the ideal product like the actual products has been successfully extended to other methodologies than intensity scales such as CATA (Ares, Dauber, Fernández, Giménez, & Varela, 2014; Ares et al., 2017; Ares, Varela, Rado, & Giménez, 2011; Bruzzone et al., 2015), Projective Mapping (Ares et al., 2011) and Pairwise Comparison (Brard & Lê, 2016). These studies suggest that characterizing the ideal product is relevant even when it is not performed using intensity measurements of each descriptor and when only a single ideal product is considered for each consumer.

Until now, most of the existing methodologies that aim at investigating drivers of liking and characterizing the ideal product are based on a pre-established list of descriptors, which comes with several limitations. The list is tedious to establish and represents a critical aspect for the relevance of the collected data as it may affect the results of the study (Ares et al., 2013). The list raises consumers' awareness on descriptors they would not think about otherwise (Coulon-Leroy, Symoneaux, Lawrence, Mehinagic, & Maitre, 2017; Kim, Hopkinson, van Hout, & Lee, 2017; Krosnick, 1999). Since the list contains only a limited number of descriptors, it could result in a loss of information and the collected data can be biased by the dumping effect (Krosnick, 1999; Varela et al., 2018). When used in a CATA task, the list likely leads to an acquiescence bias (Callegaro, Murakami, Tepman, & Henderson, 2015; Kim et al., 2017; Krosnick, 1999), which encourages consumers to check the proposed descriptors.

Luc, Lê, and Philippe (2020) took a step forward in the characterization of the ideal product without the use of a pre-established list of descriptors by proposing the so-called Free JAR profiling. In Free JAR profiling, consumers are instructed to describe a set of products using free descriptions constrained to a JAR syntax. In Free JAR profiling, the ideal product is not directly characterized since its characteristics are derived from the Free JAR descriptions of the actual products. This can result in some loss and/or some misleading information regarding the ideal product if the actual products are not carefully chosen.

Free-Comment (FC) (ten Kleij & Musters, 2003), where consumers are instructed to describe the products using their own terms into free descriptions without syntax constraint, appears as a natural alternative to identify drivers of liking and to characterize the ideal product avoiding

the limitations from the existing methodologies. Accordingly, first, the present paper investigates the relevance of FC sensory data to be linked to hedonic data with the final aim of identifying drivers of liking. Second, the Ideal-Free-Comment (IFC) method is introduced and its ability to provide a relevant characterization of the ideal product is investigated. In IFC, consumers are instructed to describe actual products and then their ideal product thanks to FC. In comparison to Free JAR profiling, IFC renders the characterization of the ideal product as independent as possible from the characterization of the actual products with the same benefit of not restricting the sensory characterizations to a pre-established list of descriptors. The final objective was to investigate whether drivers of liking and the ideal product provide consistent, and eventually complementary, information.

2. Material and methods

2.1. Participants

483 consumers from 7 French cities (Agen, Angers, Bourg en Bresse, Caen, Dijon, La Rochelle, Strasbourg) were recruited by technical centers from the ACTIA network and by the SensoStat Company. Among these consumers, 58% were females, 19% were between 18 and 30 years old, 47% were between 31 and 51 years old and 34% were more than 51 years old. They were selected as being consumers of cooked ham at least once every two weeks and were informed that they should purchase and evaluate a minimum of 4 different hams among a provided list of 30 hams widely available on the French market. Compensation for their participation was 2.5 € for each different evaluated product and no additional compensation was given to those who evaluated more than 12 different products.

2.2. Products

A list of 30 cooked hams of the French Market was selected to span the variability of fat and salt contents observed in this market. This sample was restricted to hams without rind and excluded smoked, braised, spit-roasted, and flavored hams.

2.3. Data acquisition

2.3.1. General procedure

The consumers purchased the products they evaluated and performed the evaluations at home. Each product they evaluated had to be one of the 30 products belonging to the proposed list.

An email was sent to the consumers to invite them to connect to TimeSens© (INRAE, Dijon, France) each time they evaluated a product. At each connection, the consumers had to type the European Article Numbering (*EAN*) of the ham they purchased. The consumers could not start the evaluation of a product they already evaluated as this was verified thanks to the EAN. To ensure they bought the product, they had to take a picture of the package, before and after opening. The study lasted 13 weeks and consumers could purchase hams whenever they decided but they were restricted to a maximum of one evaluation per day. Despite consumers were instructed to evaluate a minimum of 4 different hams and were compensated up to 12 ones, some of them evaluated less than 4 and others more than 12. Consequently, the number of hams evaluated by each consumer actually ranged between 1 and 14 (mean = 5.71, sd = 2.47) resulting in a total of 2758 evaluations. The data from consumers not respecting instructions were kept, as every information is good to take. The number of evaluations by ham ranged between 8 and 263 (mean = 91.93, sd = 63.38).

2.3.2. Sensory and hedonic characterization of the actual products

For each evaluated product, it was recalled to consumers to evaluate and consume the product on its own without extra food. They first performed an FC task by sensory modality in the following order: visual aspect, texture in mouth, and flavor. For each sensory modality, the following instructions were given to the consumers:

- Visual aspect: “Please describe the visual aspect of this ham”
- Texture in mouth: “Please describe the texture in mouth of this ham”
- Flavor: “Please describe the taste of this ham”

Right after the FC task, the consumers rated their liking of the product using a 0-10 VAS scale. Finally, the consumers had to provide their perception of the salt level, the fat level, the tenderness, and the color intensity of the product using 5-points Just-About-Right (JAR) scales.

After the sensory evaluation of the product, the consumers answered a few questions concerning their motivations for having purchased this product.

2.3.3. Sensory characterization of the ideal product

When the consumers decided to stop purchasing and evaluating products, they connected to TimeSens© and selected the corresponding option. This led them to answer a final questionnaire. In this questionnaire, they had to describe their ideal product using FC

descriptions according to the same three sensory modalities used to describe the actual products. For each sensory modality, the following instructions were given to the consumers:

- Visual aspect: “Please describe the visual aspect of an ideal ham in your opinion”
- Texture in mouth: “Please describe the texture in mouth of an ideal ham in your opinion”
- Flavor: “Please describe the taste of an ideal ham in your opinion”.

Some consumers did not answer the final questionnaire, resulting in a final number of 415 evaluations for the ideal product.

2.4. Data analyses

All FC data treatments and analyses were performed using R 4.0.2 (R Core Team, 2020). The lexicon provided with IRaMuTeQ© (Ratinaud, 2014) software was used for lemmatization and part-of-speech tagging.

Since the focus is on IFC, JAR scales were not analyzed in this paper.

2.4.1. FC data treatment

2.4.1.1. FC descriptions of the actual products

As FC descriptions were collected in French, all subsequent treatments were performed in French. The descriptors resulting from the treatments were then translated into English for the present paper. The English-French correspondence of the descriptors can be found in the appendix.

The FC datasets from each of the three sensory modalities (visual aspect, texture in mouth, and flavor) were treated separately with the method described in (Mahieu, Visalli, Thomas, & Schlich, 2020) and summarized thereafter. The FC descriptions of the ideal product were not involved in this process.

The descriptions were first cleaned, lemmatized, and filtered. Then, the descriptors with similar meanings were grouped into latent-descriptors relying on an ascendant hierarchical classification.

Among all the descriptors and latent-descriptors, only those mentioned throughout at least 5% of the evaluations of at least one product were retained for further analysis.

Finally, the descriptors were cross-tabulated with the consumers and the products indicating whether each descriptor was cited in the corresponding evaluation or not.

2.4.1.2. FC descriptions of the ideal product

The FC descriptions of the ideal product were treated the same manner as the FC descriptions of the actual products. They were cleaned, lemmatized, and filtered using the same filters that those used for the actual products, and the same descriptor groupings were applied. Some additional descriptors not mentioned for the actual products appeared in the descriptions of the ideal product. However, these additional descriptors were not mentioned by at least 5% of the consumers that described their ideal product and they were thus not retained for further analyses. Finally, the descriptors were cross-tabulated with the consumers indicating whether each descriptor was cited by the corresponding consumer in its description of the ideal product or not.

2.4.2. Panel level

2.4.2.1. Drivers of liking

The liking scores were regressed against the consumer factor, the product factor, and the descriptor factors using a mixed linear model fitted on all evaluations. Each descriptor factor had two levels: absence or presence, the absence level being the reference one. The descriptor factors and the product factor were considered as fixed while the consumer factor was considered as random. The regression loading of each descriptor was considered as an estimate of its impact on liking scores. Confidence intervals (α = 5%) for the regression loadings were computed using the Satterthwaite approximation (Giesbrecht & Burns, 1985; Hrong-Tai Fai & Cornelius, 1996; Satterthwaite, 1946).

2.4.2.2. Ideal product

The proportion of citations of each mentioned descriptor in the FC descriptions of the ideal product were computed. Confidence intervals (α = 5%) for these proportions were computed based on bootstrap resamplings of the consumers (1000 simulations). Descriptors significantly more frequently cited for the ideal product relatively to the actual products were investigated using multiple-response hypergeometric tests (Mahieu, Schlich, Visalli, & Cardot, 2021) with a one-sided greater alternative hypothesis (α = 5%). For these tests, the random hypergeometric samplings to estimate the null distribution were performed using the FC descriptions of the actual products provided by the consumers having described their ideal product.

For each of the three sensory modalities, a multiple-response Correspondence Analysis (MR-CA) (Mahieu et al., 2021) was performed based on the descriptor citation proportions for the

actual products. The ideal product was projected as a supplementary observation (based on its own descriptor citation proportions) into the sensory space depicted by the actual products. Confidence ellipse (α = 5%) for the ideal product coordinates was build based on bootstrap resampling of the consumers (1000 simulations). Finally, the vector of mean liking scores of the actual products was projected as a supplementary variable into the sensory space by computing its weighted correlation coefficient with the MR-CA axes and using the same weight as the MR-CA. This was performed to link the mean liking scores to the position of the ideal product.

2.4.3. Consumer segments

2.4.3.1. Segmentation of the consumers based on their ideal product

The consumers were segmented based on their FC descriptions of the ideal product considering the three sensory modalities and using a mixture-model-based clustering for nominal data (Linzer & Lewis, 2011). The model assumes the data coming from a finite mixture of K class-conditional probability distributions. The mixing proportions and the class-conditional probability distributions are estimated by maximizing the log-likelihood of the model using the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). The models ranging from K = 1 class to K = 10 classes were built. The "best" model was selected as the one having the lowest mean of its AIC (Akaike, 1974) and BIC (Schwarz, 1978). This resulted in retaining the two-class model. Finally, each consumer was affected to a class using a *maximum a posteriori* (MAP) decision rule. This resulted in two segments respectively composed of 351 (G1) and 64 (G2) consumers.

2.4.3.2. Characterization of each segment of consumers

Potential differences between the two segments in terms of gender repartition and age group repartition were investigated using a chi-square test (α = 5%). Potential differences between the two segments in terms of average frequency of consumption of cooked hams by month were investigated using a bilateral t-test (α = 5%).

2.4.3.3. Ideal product of each segment of consumers

The same computations as presented in section 2.5.1.2 were performed within each segment.

2.4.3.4. Drivers of liking of each segment of consumers

The same computations as presented in section 2.5.1.1 were performed within each segment. The drivers of liking of each segment were investigated to be compared to the ideal product of the corresponding segment.

3. Results

3.1. Panel level

3.1.1. Drivers of liking

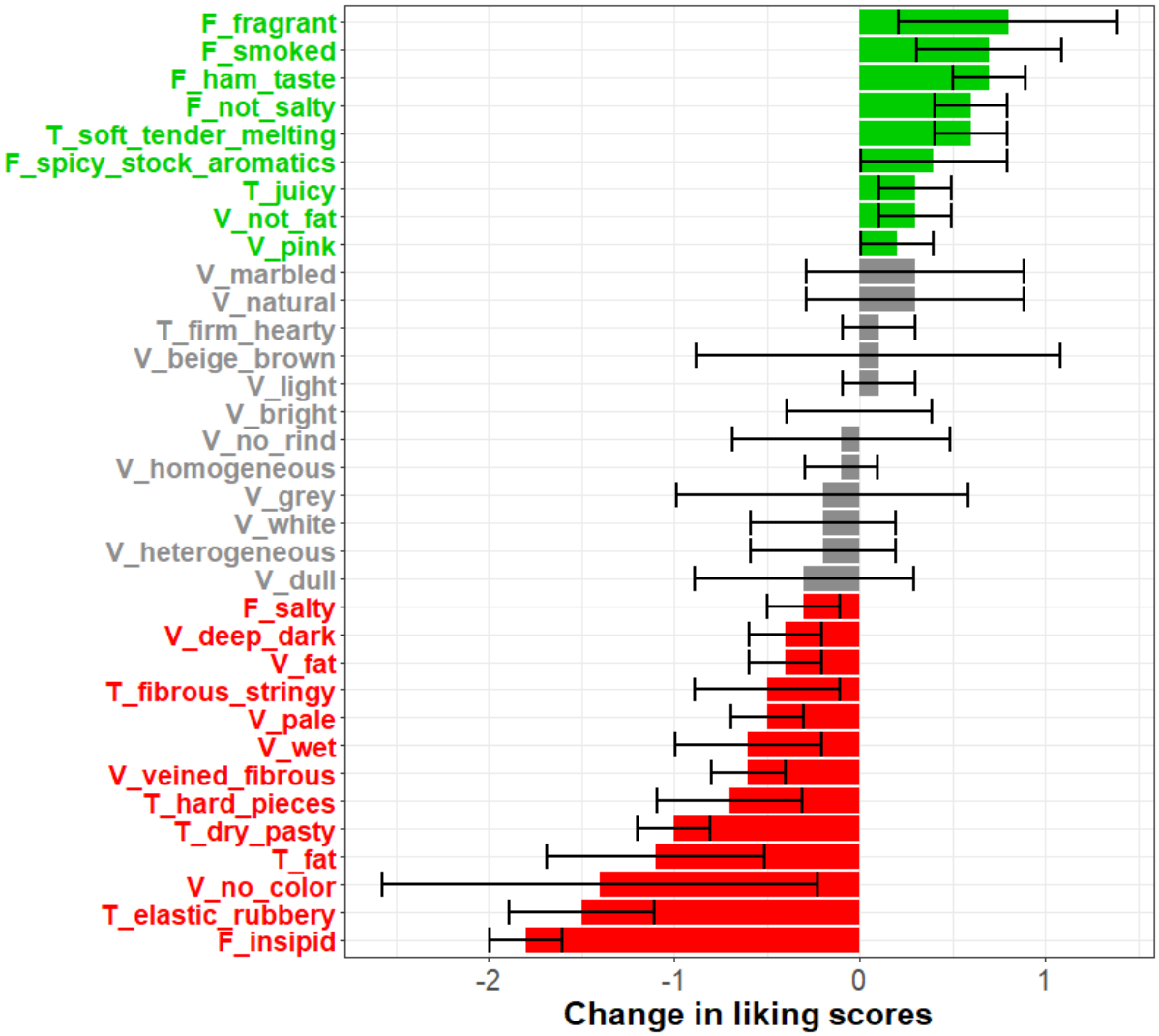


Fig. 1: Regression loadings of each descriptor with their respective confidence intervals (α = 5%). V stands for the visual descriptors, T stands for the texture in mouth descriptors and F stands for the flavor descriptors. Green (resp. red) bars represent significant (α = 5%) positive (resp. negative) drivers of liking.

Fig. 1 shows that the identified drivers of liking make sense from a sensory point of view. The negatively connoted descriptors (e.g. *F_insipid*, *T_elastic_rubbery*, etc.) were diagnosed as negative drivers of liking. On the contrary, the positively connoted descriptors (e.g *F_fragrant*, *T_soft_tender_melting*, etc.) were diagnosed as positive drivers of liking. Some less trivial

information is also shown in Fig. 1. For example, observing *F_not_salty* as a driver of liking and *F_salty* as a driver of disliking can be useful information, especially in a nutritional context. It appears in Fig. 1 that flavor impacted more liking than the texture in mouth which itself impacted more liking than the visual aspect. Finally, it can be seen in Fig. 1 that there were more drivers of disliking than drivers of liking. Also, drivers of disliking had more impact on liking scores in absolute value than drivers of liking.

3.1.2. Ideal product

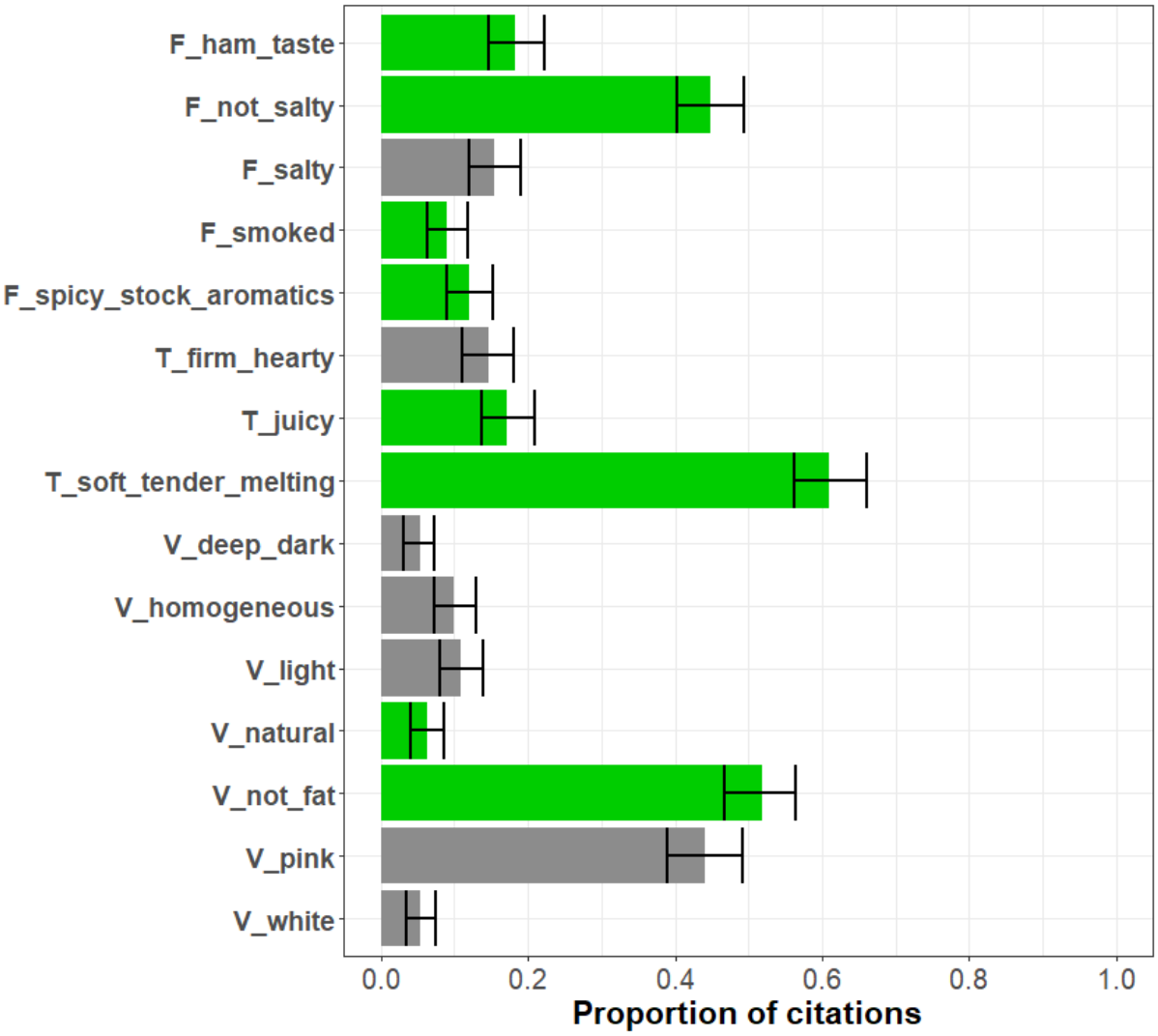


Fig. 2: Proportions of citation of descriptors mentioned in the FC descriptions of the ideal product with their respective confidence intervals (α = 5%). Green bars represent descriptors significantly more frequently cited for the ideal product relatively to the actual products (multiple-response hypergeometric test, α = 5%).

Fig. 2 shows that the mentioned descriptors in the FC descriptions of the ideal product were relevant as no negatively connoted descriptors were mentioned, which confirms that the consumers understood the concept of describing their ideal product. Some characteristics appeared very important to be found in the ideal product: *V_not_fat*, *V_pink*, *V_soft_tender*,

F_not_salty. The descriptors significantly more frequently cited for the ideal product relatively to the actual products were consistent with the identified drivers of liking. However, some differences can still be noticed. One descriptor significantly associated with the ideal product was not identified as a driver of liking: a "natural" visual appearance (*V_natural*). On the contrary, *F_fragrant* identified as a driver of liking was not cited in the FC descriptions of the ideal product. Finally, some opposite descriptors (e.g. *F_salty* vs. *F_not_salty*) were mentioned in FC descriptions of the ideal product which justifies investigating if consumer segments exist (see Section 3.2).

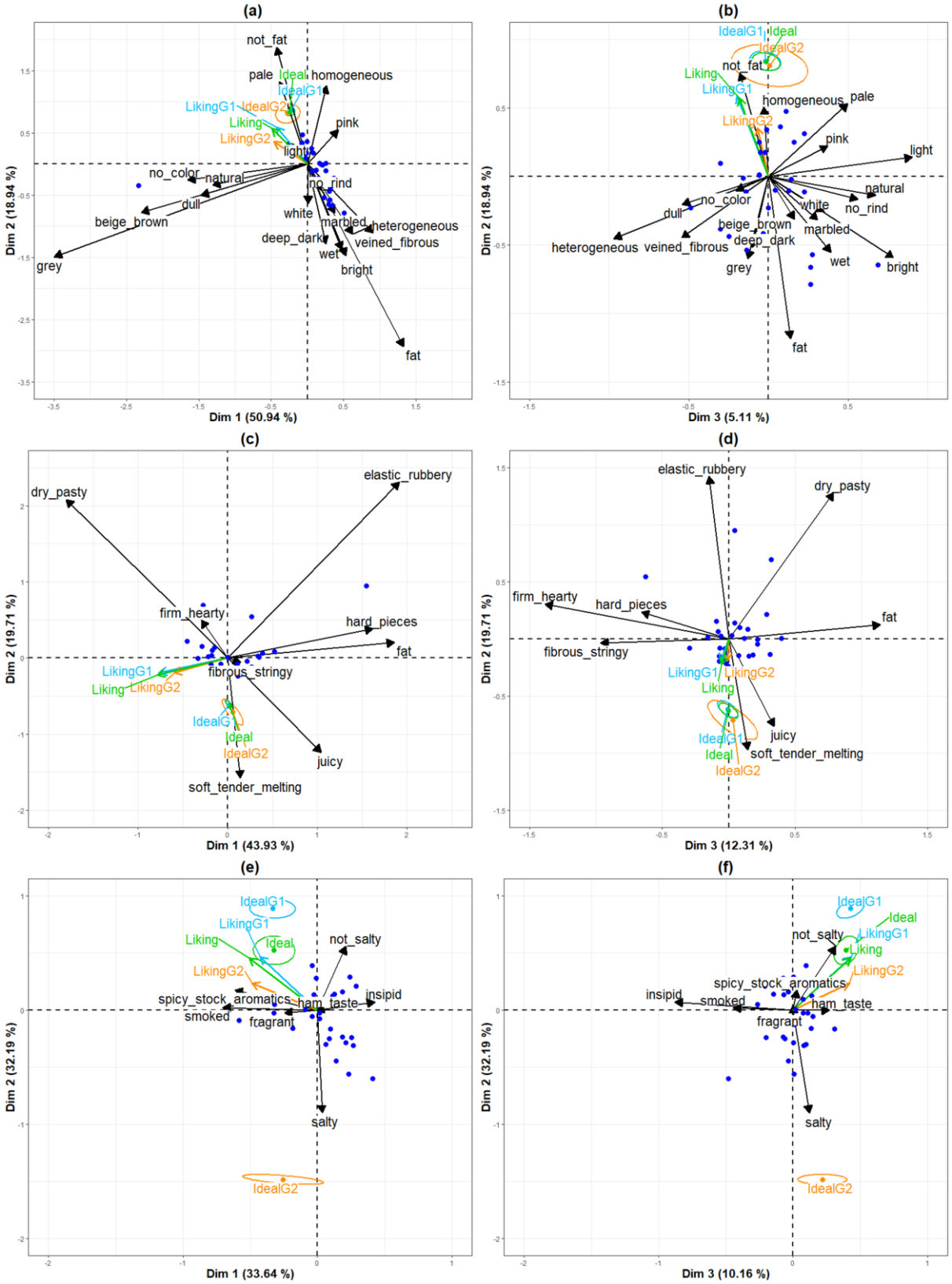


Fig. 3: Biplot from multiple-response Correspondence Analysis with the panel (Ideal) and by segment (IdealG1 and IdealG2) ideal products (projected as supplementary observation), their confidence ellipse (α = 5%) and the mean panel (Liking) and by segment (LikingG1 and LikingG2) liking scores (projected as supplementary variable): (a) axes 1-2 visual aspect, (b) axes 3-2 visual aspect, (c) axes 1-2 texture in mouth, (d) axes 3-2 texture in mouth, (e) axes 1-2 flavor, (f) axes 3-2 flavor. Blue points are the actual products (unlabeled for sake of readability). Weighted correlation values of liking scores can be read thanks to the axes ticks.

For the three sensory modalities, Fig. 3 (note that *IdealG1*, *IdealG2*, *LikingG1* and *LikingG2* refer to a subsequent segmentation discussed later in section 3.2) shows that the ideal product achieved the most extreme coordinates in the direction of the liking among all the products and lied in a region of the sensory space that none of the actual products reached. The first point confirms that the consumers understood the concept of describing their ideal product and suggests that they provided ideal product descriptions consistent with their liking scores. The second point suggests that none of the actual products was ideal and that gathering descriptions of the ideal product can provide relevant information. It is worth noticing that even if the ideal product lied in a particular region of the sensory space, it was not the most distant product from the average. This statement is true for the three sensory modalities and suggests that the ideal product might be realistic. Interestingly, the confidence ellipse of the ideal product was larger for the flavor modality than for the two other sensory modalities. This is likely because consumers were more consensual in describing their ideal product regarding visual aspect and texture in mouth than regarding flavor and reinforces that investigating if consumer segments exist might be relevant.

Fig. 3 also suggests that the flavor modality is the most important regarding hedonic appreciation. This is further confirmed by the average absolute weighted correlation of the mean liking scores with the whole sensory axes: 0.139 for visual aspect, 0.201 for texture in mouth, and 0.261 for flavor. Finally, it is interesting to notice that this ranking of the sensory modalities regarding the link between their sensory axes and the mean liking scores is the same as the ranking observed for the drivers of liking regarding the impact of each sensory modality on the liking scores.

3.2. Consumer segments

3.2.1. Characterization of each segment of consumers

The two segment were not statistically different regarding their gender repartition ($Chi^2 = 0.074$, $df = 1$, $p = 0.7857$), their age group repartition ($Chi^2 = 2.771$, $df = 2$, $p = 0.2502$), and their average frequency of consumption of cooked hams by month ($t = -0.5802$, $df = 374$, $p = 0.5621$).

3.2.2. Ideal product of each segment of consumers

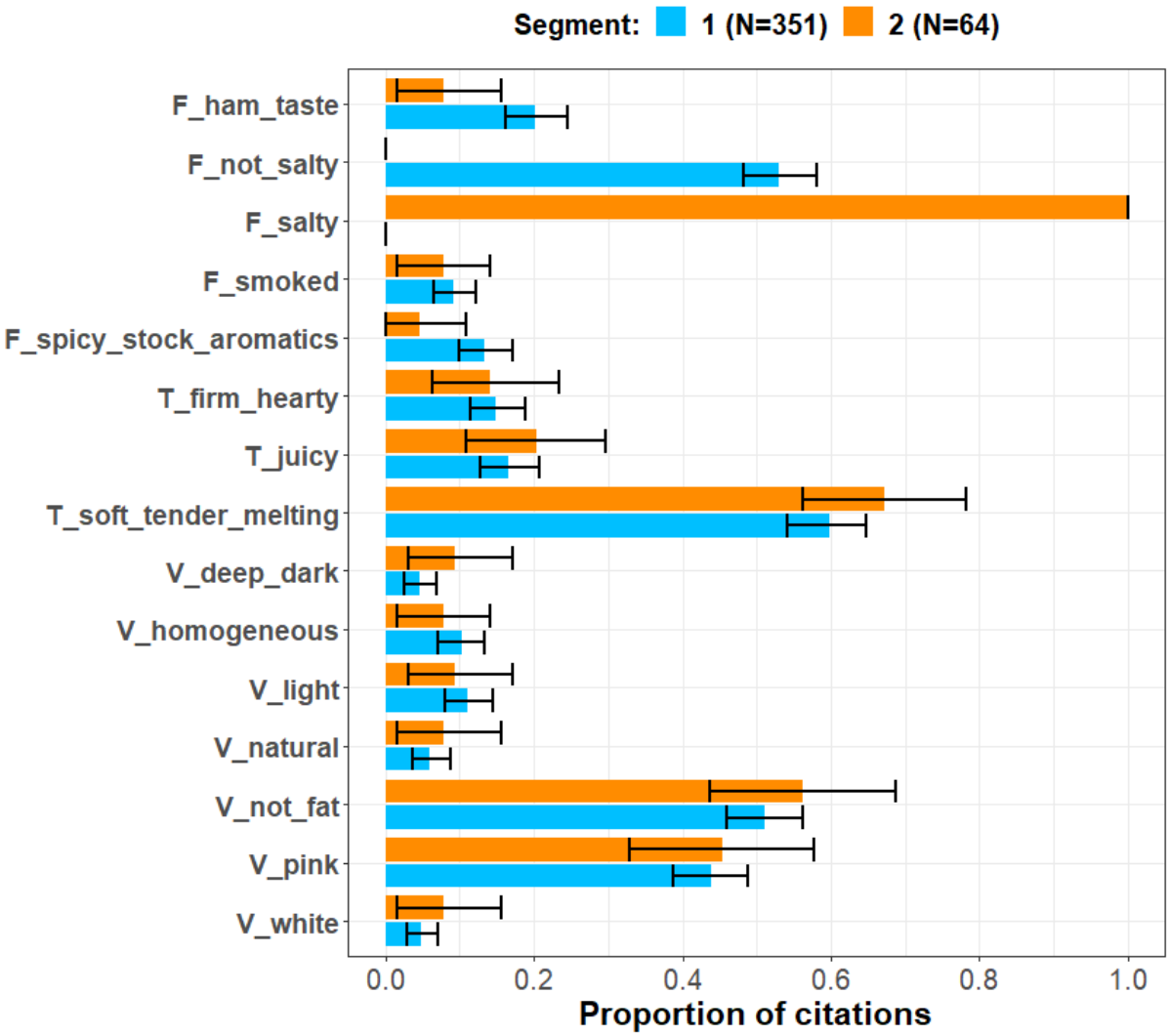


Fig. 4: Proportion of citations of descriptors mentioned in the FC descriptions of the ideal product within each segment with their respective confidence intervals (α = 5%).

Fig. 4 shows that the two segments of consumers are interpretable. The ideal products of the two segments mainly differed regarding their flavor. The ideal product of G1 was described as *F_not_salty* approximately half of the time while it was never described as *F_salty*. On the contrary, the ideal product of G2 was always described as *F_salty* while it was never described as *F_not_salty*. This suggests that two types of consumers exist. Those that would like their ideal product not to be salty and those that would like their ideal product to be salty, the "salty lovers" being fewer (≈ 15% of the consumers) than the others. Other smaller differences can be noticed between the ideal products of the two segments: the ideal product of G1 was more often described as *F_ham_taste* and *F_spicy_stocks_aromatics* than the one of G2.

Fig. 3 confirms the results from Fig. 4: the ideal products of the two segments differed regarding their flavor but neither their texture in mouth nor their visual aspect. Not surprisingly, regarding the flavor modality, the two ideal products were opposed on the second dimension, which was

a gradient of saltiness. Fig. 3 shows that the ideal product of G1 is very close and thus similar to that of the panel. This makes sense since G1 represents an overwhelming majority as compared to G2. Regarding, the mean liking scores, the two segments appeared to have a similar pattern, close to that of the panel. G1 seemed more consistent than G2 because its ideal product is located farther away in the direction of its mean liking scores for the flavor modality.

3.2.3. Drivers of liking of each segment of consumers

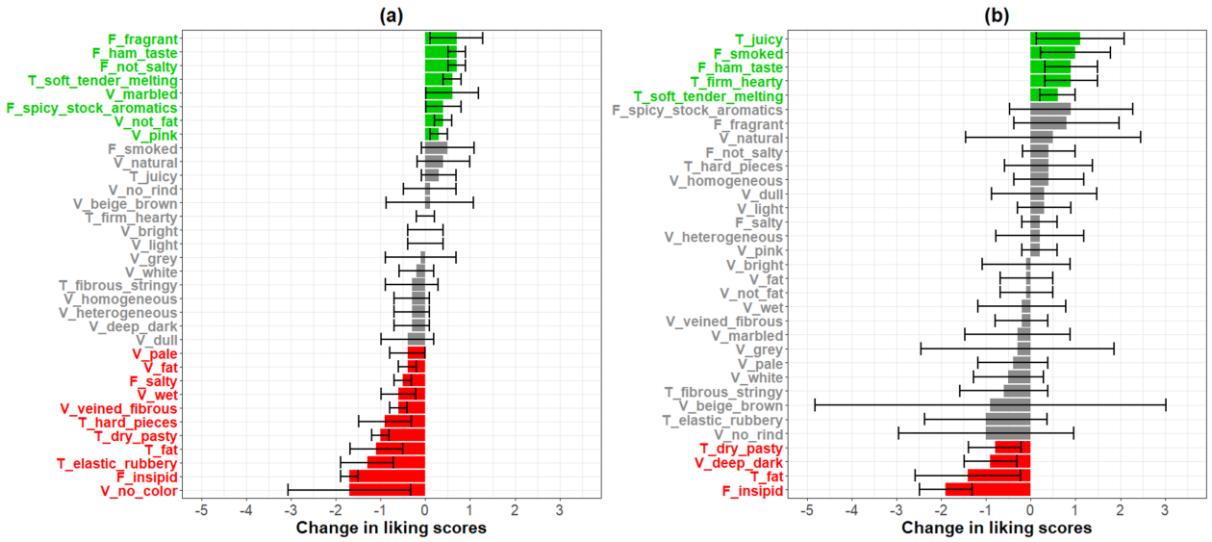


Fig. 5: Regression loadings of each descriptor with their respective confidence intervals (α = 5%) for the two segments of consumers: (a) G1 (N = 351) and (b) G2 (N = 64). V stands for the visual descriptors, T stands for the texture in mouth descriptors and F stands for the flavor descriptors. Green (resp. red) bars represent significant (α = 5%) positive (resp. negative) drivers of liking.

Fig. 5 shows that the drivers of liking of each segment were only partially consistent with their corresponding ideal product. Regarding saltiness, which was the main difference between the two ideal products, the drivers of liking of G1 were perfectly consistent with its ideal product: *F_not_salty* was a driver of liking and *F_salty* was a driver of disliking. For G2, the loading of *F_salty* was positive as opposed to this same loading for G1, but not significant. The loading of *F_not_salty* was also not significant but it was positive and higher than that of *F_salty*. This reinforces the evoked doubt (Fig. 3) on the consistency of G2. Regarding *F_ham_taste* and *F_spicy_stocks_aromatics*, which were the other main differences between the two ideal products, *F_ham_taste* was a driver of liking for the two segments and with the same intensity and *F_spicy_stocks_aromatics* was a driver of liking for G1 and not G2. However, this difference between G1 and G2 might be due to the different number of consumers in the two segments which led to the confidence intervals of G2 being larger than for G1. The fact that the loading of *F_spicy_stocks_aromatics* was higher for G2 than for G1 reinforces this line of reasoning. Overall, the main differences between the two ideal products were only moderately

recovered by comparing the drivers of liking of each segment. However, regarding the most important difference, which was the level of saltiness, G1 had drivers of liking consistent with its ideal product and a trend of consistency existed for G2 since its loading for *F_salty* was positive as opposed to G1.

4. Discussion

4.1. Drivers of liking vs. ideal product

The ideal product and the drivers of liking are different approaches that have their benefits and drawbacks. The drivers of liking are implicit and thus not subject to cognitive and attitudinal bias unlike ideal product descriptions (Li, Hayes, & Ziegler, 2015). However, drivers of liking depend on the actual product space. This constraint could result in some loss and/or some misleading information if too many sensory characteristics are confused and/or not well represented by the actual product space. Since the ideal product does not depend directly on the actual product space, it enables exploring a larger sensory space than that depicted by the actual products (Worch, Crine, Gruel, & Lê, 2014).

Overall, the ideal product and the drivers of liking should be considered complementary rather than competitors: they reinforce and validate each other. Drivers of liking which are significantly and frequently associated with the ideal product are definitely important characteristics regarding appreciation. In the specific context of FC, they are even more complimentary since some obvious and logical characteristics (e.g. *F_fragrant* in this study) may not be mentioned in the descriptions of the ideal product, as they are essential and natural. On the contrary, some characteristics confused and/or rarely present in the actual products (e.g. *V_natural* in this study) can be caught only thanks to the ideal product characterization.

In this study, drivers of liking and the panel's average ideal product provided information in agreement with each other. This suggests that this information can be used from a product development point of view. Especially, including less salt in the manufacturing process of the cooked hams would be beneficial from a nutritional point of view and could possibly increase hedonic appreciation, but certainly not decrease it.

4.2. Panel level vs. consumer segments for the ideal product

To the best of our knowledge, only one study previously proposed to segment the consumers based on their ideal product (Chan, Kwong, & Hu, 2012). Segmenting the consumers based on

their ideal product makes sense only in two situations. The first one is when opposite descriptors (e.g. *salty* vs. *not_salty*) are used in individual ideal product descriptions. The second case is when the description of the ideal product is highly variable among consumers. To determine if segmenting the consumers is relevant, and when it is, the number of segments to consider should be determined using objective criterions. Depending on the strategy of clustering adopted, different criterions exist. When mixture models are used, as in this study, information criterions such as AIC (Akaike, 1974) and BIC (Schwarz, 1978) can be used. When hierarchical clustering and/or *k*-means algorithm are used, quality of clustering indexes such as the Silhouette index (Rousseeuw, 1987) and the Gap statistic (Tibshirani, Walther, & Hastie, 2001) can be used.

Even when segmenting the consumers based on their ideal product appears relevant from both a qualitative and a statistical point of view, checking the consistency of each segment is important (Brard & Lê, 2016; Worch et al., 2014; Worch, Lê, Punter, & Pagès, 2012a, 2012b). If the ideal product of one or more segments does not make sense regarding their drivers of liking, segmenting the consumers is questionable. Similarly, when the segments share common drivers of liking but have a different ideal product, segmenting is questionable. In this context, to better understand the differences between the ideal products of each segment, using mapping techniques (e.g. factorial analyses) and absolute measurements (e.g. probabilities of citations) are useful and should be used conjointly. Further, considering that some consumers could eventually provide ideal product descriptions based on non-sensory criteria (e.g. health) (Worch et al., 2013) could help understanding some non-consistent segments. Indeed some consumers could like sweet products but their ideal product could be described as not sweet because they are diabetics for example. However, since the ideal descriptions are instructed to be provided based on the sensory perception (visual aspect, texture in mouth and flavor in this study) this is unlikely to occur.

If different segments of consumers are identified, but one or some of them are of a too-small size, then one should not consider the segmentation (Worch et al., 2012a, 2012b).

In the present paper, G1 highly dominated G2 in terms of size. Further, the consistency of G2 was highly questionable, and G1 and G2 had no clear difference in their drivers of liking except maybe on the level of saltiness. This suggests that for this paper, the analyses performed at the panel level considering a single ideal product are likely the most relevant. Alternatively, as suggested by (Worch et al., 2012a, 2012b), the ideal product descriptions coming from the consumers of G2 could be dropped from the analysis by considering only those from G1.

4.3. Limitations

A first limitation comes from the uncommon data collection procedure of this study. Indeed, to the best of our knowledge, it is the first time that sensory and hedonic data are gathered from consumers purchasing the products they evaluate, which resulted in unbalanced data for the actual products. This uncommon procedure does not appear to be a limitation as the data make sense. However, it worth emphasizing that the liking scores of the actual products may have been overestimated. Indeed, because consumers selected the products they evaluated, some of them may have selected products they usually purchase and like. Knowing that 20% of the evaluations among the 2758 ones were performed on usually purchased hams and that an average overall liking score of 6.35 (all products combined) was observed, the previous assertion could be at least partly verified. However, other strategies of selection from the consumers may have occurred such as selecting less expensive ones to maximize income from compensations, testing more expensive ones as they were partly refunded by the compensations, or selecting hams based on their labels and/or allegations. These other strategies, considered together with the requirement that, for being compensated, the consumers had to evaluate at least 4 different hams from the list, are the most likely explanations to the fact that most of the hams belonging to the list were evaluated a fair number of times, thus limiting the liking overestimation. Another point that worth emphasizing is that, since the consumers selected their evaluated products, they may have restricted the product space and with that, the range of encountered sensory characteristics, which may have affected the ideal product descriptions provided after the evaluations of actual products. Indeed, consumers likely defined what they like and dislike based on the evaluations of actual products. Depending on the practitioners' aims, if gathering less "informed" ideal product descriptions is of interest, consumers could be instructed to provide them before evaluations of actual products but this could inversely affect actual products descriptions. Anyway, investigating the method presented in this paper with a more "conventional" experimental procedure might be an interesting direction for some future research. In particular, comparing the consumer segments resulting from a segmentation on either ideal product data or liking data would be of great interest. Segmenting consumers based on liking data was not performed in this study because of the uncommon experimental design that resulted in a "product by consumer" matrix of liking scores having 81% of missing data.

A second limitation comes from the IFC method and the data analysis procedure proposed in this study. More specifically, if some descriptors not present in the FC descriptions of the actual products are mentioned in the FC descriptions of the ideal product, the projection of the ideal

product into the sensory space depicted by the actual products can only be performed on basis of the descriptors shared by the actual and the ideal product descriptions. However, this is not a major limitation since the aim of this projection is to investigate the position of the ideal product relative to the actual products, which make sense to be performed on the same set of descriptors. All the other analyses presented in this study can be performed equivalently with additional descriptors for the ideal product as compared to the actual products. Finally, it has to be mentioned that if this situation occurs, it is a nice argument in favor of IFC since no other existing method can investigate the hedonic importance of descriptors not present within the actual product space.

5. Conclusion

The paper proposes to use Free-Comment (FC) sensory data to be used in the well-established link between sensory and hedonic data. Further, it introduced a new methodology called Ideal-Free-Comment (IFC) where consumers are instructed to describe actual products and then their ideal product thanks to FC. This enables investigating drivers of liking and characterizing the ideal product without the use of a pre-established list of descriptors, which *de facto* avoids inherent limitations to any pre-established list. Further, since the characterization of the ideal product is directly performed, it does not depend on the actual product space, and the hedonic importance of descriptors confused and/or rarely present in the actual products can thus be investigated. Identification of drivers of liking based on FC data and IFC were used on cooked hams with consumers purchasing the products they evaluated at home and it showed relevant results. Drivers of liking based on FC data and IFC provide sensory analysts with new complementary tools to understand consumers' hedonic appreciation without the use of a pre-established list of descriptors.

Appendix: English-French correspondence of the descriptors

Sensory modality	English	French
Visual aspect	beige_brown	beige_marron
	bright	brillant
	deep_dark	foncé_sombre
	dull	terne
	fat	gras

	grey	gris
	heterogeneous	hétérogène
	homogeneous	homegène
	light	clair
	marbled	marbré
	natural	naturel
	no_color	sans_couleur
	no_rind	sans_couenne
	not_fat	pas_gras
	pale	pâle
	pink	rose
	veined_fibrous	nevuré_fibreux
	wet	humide
	white	blanc
Texture in mouth	dry_pasty	sec_pâteux
	elastic_rubbery	élastique_caoutchouteux
	fat	gras
	fibrous_stringy	fibreux_filandreux
	firm_hearty	ferme_consistant
	hard_pieces	morceaux_durs
	juicy	juteux
	soft_tender_melting	moelleux_tendre_fondant
Flavor	fragrant	parfumé
	ham_taste	goût_de_jambon
	insipid	fade
	not_salty	pas_salé
	salty	salé
	smoked	fumé
	spicy_stock_aromatics	épicé_bouillon_aromates

Acknowledgments

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

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.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Ares, G., Dauber, C., Fernández, E., Giménez, A., & Varela, P. (2014). Penalty analysis based on CATA questions to identify drivers of liking and directions for product reformulation. *Food Quality and Preference, 32*, 65-76.

Ares, G., de Andrade, J. C., Antúnez, L., Alcaire, F., Swaney-Stueve, M., Gordon, S., et al. (2017). Hedonic product optimisation: CATA questions as alternatives to JAR scales. *Food Quality and Preference, 55*, 67-78.

Ares, G., Jaeger, S. R., Bava, C. M., Chheang, S. L., Jin, D., Gimenez, A., et al. (2013). CATA questions for sensory product characterization: Raising awareness of biases. *Food Quality and Preference, 30*(2), 114-127.

Ares, G., & Varela, P. (2017). Trained vs. consumer panels for analytical testing: Fueling a long lasting debate in the field. *Food Quality and Preference, 61*, 79-86.

Ares, G., Varela, P., Rado, G., & Giménez, A. (2011). Identifying ideal products using three different consumer profiling methodologies. Comparison with external preference mapping. *Food Quality and Preference, 22*(6), 581-591.

Brard, M., & Lê, S. (2016). The Ideal Pair Method, an Alternative to the Ideal Profile Method Based on Pairwise Comparisons: Application to a Panel of Children. *Journal of Sensory Studies, 31*(4), 306-313.

Bruzzone, F., Vidal, L., Antúnez, L., Giménez, A., Deliza, R., & Ares, G. (2015). Comparison of intensity scales and CATA questions in new product development: Sensory characterisation and directions for product reformulation of milk desserts. *Food Quality and Preference, 44*, 183-193.

Callegaro, M., Murakami, M. H., Tepman, Z., & Henderson, V. (2015). Yes-no answers versus check-all in self-administered modes. *International Journal of Market Research, 57*, 203-223.

Carroll, J. D. (1972). Individual Differences and Multidimensional Scaling. In, *R. N. Shepard, A. K. Romney, & Si Nerloves (Eds.), Multidimensional scaling: Theory and applications in the behavioral sciences*. New York: Academic Press.

Chan, K. Y., Kwong, C. K., & Hu, B. Q. (2012). Market segmentation and ideal point identification for new product design using fuzzy data compression and fuzzy clustering methods. *Applied Soft Computing, 12*(4), 1371-1378.

Coulon-Leroy, C., Symoneaux, R., Lawrence, G., Mehinagic, E., & Maitre, I. (2017). Mixed Profiling: A new tool of sensory analysis in a professional context. Application to wines. *Food Quality and Preference, 57*, 8-16.

Danzart, M. (2009). *SSHA 3eme (Ed.), Evaluation sensorielle. Manuel méthodologique*. Paris: Lavoisier.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1-38.

Dooley, L., Lee, Y.-s., & Meullenet, J.-F. (2010). The application of check-all-that-apply (CATA) consumer profiling to preference mapping of vanilla ice cream and its comparison to classical external preference mapping. *Food Quality and Preference, 21*(4), 394-401.

Escofier, B., & Pagès, J. (1994). Multiple factor analysis (AFMULT package). *Computational Statistics & Data Analysis, 18*(1), 121-140.

Faye, P., Brémaud, D., Teillet, E., Courcoux, P., Giboreau, A., & Nicod, H. (2006). An alternative to external preference mapping based on consumer perceptive mapping. *Food Quality and Preference, 17*(7), 604-614.

Giesbrecht, F. G., & Burns, J. C. (1985). Two-Stage Analysis Based on a Mixed Model: Large-Sample Asymptotic Theory and Small-Sample Simulation Results. *Biometrics, 41*(2), 477-486.

Greenhoff, K., & MacFie, H. J. H. (1994). Preference mapping in practice. In, *H. J. H. MacFie & D. M. H. Thompson (Eds.), Measurement of food preferences*. London: Blackie Academic & Professionals.

Grygorczyk, A., Lesschaeve, I., Corredig, M., & Duizer, L. (2013). Extraction of consumer texture preferences for yogurt: Comparison of the preferred attribute elicitation method to conventional profiling. *Food Quality and Preference, 27*(2), 215-222.

Hrong-Tai Fai, A., & Cornelius, P. L. (1996). Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation, 54*(4), 363-378.

Kim, I.-A., Hopkinson, A., van Hout, D., & Lee, H.-S. (2017). A novel two-step rating-based 'double-faced applicability' test. Part 1: Its performance in sample discrimination in comparison to simple one-step applicability rating. *Food Quality and Preference, 56*, 189-200.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537-567.

Lagrange, V., & Norback, J. P. (1987). Product optimization and the Acceptor Set Size. *Journal of Sensory Studies, 2*(2), 119-136.

Lawless, H. T., & Heymann, H. (1999). *Sensory evaluation of food: Principles and practices*. New York: Kluwer Academic/Plenum Publishers.

Li, B., Hayes, J. E., & Ziegler, G. R. (2015). Maximizing overall liking results in a superior product to minimizing deviations from ideal ratings: An optimization case study with coffee-flavored milk. *Food Quality and Preference, 42*, 27-36.

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R Package for Polytomous Variable Latent Class Analysis. *Journal of Statistical Software, 42*(10).

Luc, A., Lê, S., & Philippe, M. (2020). Nudging consumers for relevant data using Free JAR profiling: An application to product development. *Food Quality and Preference, 79*.

Mahieu, B., Schlich, P., Visalli, M., & Cardot, H. (2021). A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data. *Food Quality and Preference, 93*.

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

McEwan, J. A. (1996). Preference Mapping for product optimization. In, *T. Naes & E. Risvik (Eds.), Multivariate analysis of data in sensory science*. New York: Elsevier.

Meilgaard, M., Civille, G. V., & Carr, B. T. (1991). *Sensory Evaluation Techniques*. Boca Raton, FL: CRC Press.

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.

Moskowitz, H. R. (1972). Subjective ideals and sensory optimization in evaluating perceptual dimensions in food. *Journal of Applied Psychology, 56*(1), 60-66.

Popper, R. (2014). Use of Just-About-Right Scales in Consumer Research. In, *P. Varela & G. Ares (Eds), Novel techniques in sensory characterization and consumer profiling*. Boca Raton, FL (United States): CRC Press.

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

Ratinaud, P. (2014). IRaMuTeQ : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. In. France.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65.

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin, 2*(6), 110-114.

Schlich, P., & McEwan, J. A. (1992). Cartographie des préférences. Un outil statistique pour l'industrie agro-alimentaire. *Science des Aliments, 12*, 339-355.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6*(2), 461-464.

Stone, H., & Sidel, J. L. (1993). *Sensory evaluation practices*. California: Academic Press.

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.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411-423.

Valentin, D., Chollet, S., Lelièvre, M., & Abdi, H. (2012). Quick and dirty but still pretty good: a review of new descriptive methods in food science. *International Journal of Food Science & Technology, 47*(8), 1563-1578.

van Kleef, E., van Trijp, H. C. M., & Luning, P. (2006). Internal versus external preference analysis: An exploratory study on end-user evaluation. *Food Quality and Preference, 17*(5), 387-399.

van Trijp, H. C. M., Punter, P. H., Mickartz, F., & Kruithof, L. (2007). The quest for the ideal product: Comparing different methods and approaches. *Food Quality and Preference, 18*(5), 729-740.

Varela, P., Antúnez, L., Carlehög, M., Alcaire, F., Castura, J. C., Berget, I., et al. (2018). What is dominance? An exploration of the concept in TDS tests with trained assessors and consumers. *Food Quality and Preference, 64*, 72-81.

Varela, P., & Ares, G. (2012). Sensory profiling, the blurred line between sensory and consumer science. A review of novel methods for product characterization. *Food Research International, 48*(2), 893-908.

Worch, T. (2013). PrefMFA, a solution taking the best of both internal and external preference mapping techniques. *Food Quality and Preference, 30*(2), 180-191.

Worch, T., Crine, A., Gruel, A., & Lê, S. (2014). Analysis and validation of the Ideal Profile Method: Application to a skin cream study. *Food Quality and Preference, 32*, 132-144.

Worch, T., Lê, S., Punter, P., & Pagès, J. (2012a). Assessment of the consistency of ideal profiles according to non-ideal data for IPM. *Food Quality and Preference, 24*(1), 99-110.

Worch, T., Lê, S., Punter, P., & Pagès, J. (2012b). Extension of the consistency of the data obtained with the Ideal Profile Method: Would the ideal products be more liked than the tested products? *Food Quality and Preference, 26*(1), 74-80.
Worch, T., Lê, S., Punter, P., & Pagès, J. (2013). Ideal Profile Method (IPM): The ins and outs. *Food Quality and Preference, 28*(1), 45-59.

Chapter VI:

Discussion and perspectives

A. Benefits and limitations of Free-Comment

This section comes back to some of the benefits and limitations of Free-Comment (FC) mentioned in chapter I but looking at them through the prism of the studies presented in this thesis and their results.

1. Benefits

The milk chocolate study from chapter IV confirmed that lists of sensory descriptors might lead consumers to consider as applicable some descriptors they would not have thought of and mentioned without the list. This is the case of the descriptor "*cereal"* of the Check-All-That-Apply (CATA) list dedicated to the description of the flavor in mouth modality of this study. This descriptor was cited in 50% of the evaluations of one product and 19% of the entire evaluations (all products combined) while it was not even mentioned in a single description of the FC group. This confirms that list-based methods can bias the sensory characterization of the products and/or influence them by the practitioners' point of view.

The wine study from chapter IV and the dark chocolate study from chapter V confirmed the existence of the dumping effect. The wine "*Bor"* was characterized by "*Light_red*" in the CATA group while it was characterized by "*brown_tile_red_orange*" in the FC group. The closest descriptor to the tile-red color in the CATA list having been "*Light_red*" the wine "*Bor"* ended up being characterized by this descriptor while the spontaneous FC descriptions tell another story. For the Free-Comment Attack-Evolution-Finish (FC-AEF) dark chocolate study, the citation rates of "*crunchy_hard*" are highly correlated ($r = 0.926$, $p < 0.001$), based on pairs of product and period as observations, to those of "*Dry*" from the Attack-Evolution-Finish (AEF) study of Visalli et al. (2020b). The closest descriptor to the crunchy texture in the AEF list having been "*Dry*" it is likely that consumers used it to report crunchiness. These assessments again

confirm that list-based methods can bias the sensory characterization of the products and/or influence them by the practitioners' point of view.

Every use of FC-based methods of this thesis provided additional sensory descriptors strongly associated with some products compared to their respective list-based homologs. Some examples can be mentioned: "*insipid*", "*crunchy_hard*", "*spicy*", and "*powdery_mealy_granular*" in the dark chocolate study of chapter V, "*insipid*" in the milk chocolate study of chapter IV, or "*dry"* and "*short_tasting*" in the wine study of chapter IV. These descriptors are all relevant descriptive sensory information confirming that list-based can miss some information.

Study	Sensory modality or temporal period	Method	Number of descriptors	Average (± 95% confidence interval) percentage of descriptors used by one consumer all products combined	Average (± 95% confidence interval) percentage of descriptors used by one consumer per evaluation
Wine	Visual	FC	12	38.3% ± 3.8%	18.5% ± 1.2%
		CATA	8	55.0% ± 4.3%	28.2% ± 1.4%
	Olfactory	FC	14	30.4% ± 3.0%	11.5% ± 1.0%
		CATA	10	46.5% ± 4.0%	20.1% ± 1.3%
	Gustatory	FC	20	30.6% ± 3.0%	11.9% ± 0.9%
		CATA	19	58.0% ± 3.5%	25.3% ± 1.1%
Milk chocolate	Texture in mouth	FC	10	40.3% ± 2.7%	16.1% ± 0.9%
		CATA	8	71.2% ± 4.2%	26.7 ± 1.4%
	Flavor in mouth	FC	8	35.4% ± 2.9%	15.1% ± 1.1%
		CATA	6	66.4% ± 4.3%	29.5% ± 2.4%

Table 3: Descriptive statistics on citations rates for the datasets of chapter IV

Comparing the citation rates of FC to the citation rates of CATA in the studies of this thesis confirms the acquiescence bias. Indeed, as shown in Table 3, regardless of the study and the sensory modality, the consumers used fewer descriptors from the list in FC as compared to CATA. Similarly, citation rates of descriptors by evaluation were always lower in FC as compared to CATA. Since the number of descriptors between FC and CATA for a given study and sensory modality was on the same order of magnitude, these differences are very likely to be indeed attitudinal and not artifactual. A recent study (not yet published at the time this manuscript is written) from our team definitely confirms that these

differences are attitudinal. In this study, 98 consumers performed a FC on 4 dark chocolates and the resulting *a posteriori* list of sensory descriptors was used into a CATA performed by these same 98 consumers on the same 4 dark chocolates two weeks later. The average citation rate by evaluation for the CATA was approximately twice that of the FC. These assessments confirm that consumers are more easily inclined to consider a descriptor applicable when presented in a list rather than when they need to volunteer it. This attitudinal difference between FC and CATA suggests that list-based methods could be more "powerful" (easily detect present sensory characteristics) but also more "risky" (easily detect questionable sensory characteristics) than FC-based methods. Of course, the previous assertion requires for being verified that the *a priori* lists are properly established not to miss information or not to bias characterizations, which, as mentioned in chapter I, can be tedious and time-consuming.

2. Limitations

The studies of this thesis confirmed that some broad and general sensory descriptors are often present in the *a posteriori* list of sensory descriptors when using FC-based methods. The most representative and recurrent of them are descriptors referring to the two poles of the "strength" dimension of the products such as *gentle*, *slight*, *fresh*, *strong*, *intense*, *powerful*, etc. Despite these descriptors are indeed general, they still provide information and they contribute to understanding better the product structure and the sensory characteristics consumers consider as being related to the two poles of the "strength" dimension.

Despite the consumers were driven to provide only descriptive sensory information in their descriptions, every study of this thesis included hedonic information in its raw descriptions. Usually, raw descriptions containing hedonic information did not contain descriptive sensory information suggesting that when consumers cannot analytically describe their perception, they take refuge in their subjective affective perception of the product. This confirms that some consumers

encounter difficulty to analytically dissect and/or properly verbalize their perception when they are left to their own devices to do so, i.e. not guided by some propositions of sensory descriptors that may be applicable.

B. Gathering and pretreatment of Free-Comment data

1. Gathering of Free-Comment data

In every study of this thesis but these including a temporal component, consumers provided one Free-Comment (FC) description by sensory modalities of the products under investigation with neither restriction on the nature, the size and the structure of these descriptions. Based on the studies of this thesis, this FC data gathering procedure by sensory modality appears relevant as it enabled gathering rich characterization of the products not redundant across sensory modalities. Since FC data were gathered the same way in every study, the potential impact of the different reported alternatives (Ares et al., 2010; Hanaei et al., 2015; ten Kleij & Musters, 2003) was not investigated. Further, hedonic-oriented FC (Lahne et al., 2014; Luc et al., 2020; Symoneaux et al., 2012) was not considered in this thesis because only "pure" and non-oriented descriptive sensory characterization of the products was under interest. To the best of our knowledge, the different reported alternatives to gather FC data have never been directly compared. Comparing these alternatives would be an interesting topic to shade more light on their respective benefits and limitations depending on the products and the aims of the study.

Other alternatives of FC data gathering not reported in the literature nor investigated in this thesis could be mentioned. A first alternative would be to present the products with repetitions to the consumers to evaluate the repeatability of the individual descriptions and/or the reproducibility of the panel characterizations. However, this could rapidly increase the cognitive fatigue of consumers that are further expected not to be highly repeatable, as they are not calibrated. A second alternative would be to gather FC data using voice

recognition technologies. This could drive consumers to be more expressive resulting in more gathered information. However, as speech recognition libraries require terms to be contextualized to achieve good accuracy, this could also lead to gather more noise resulting in harder pretreatment and interpretation of FC data. In some practical applications, where products under interest would not enable consumers to type their descriptions easily (e.g. hand soap or cream, sportswear, etc.), voice recognition could be of paramount interest. A third alternative would be, within the same session, to first instruct consumers to provide FC descriptions, then extract relevant sensory descriptors and finally instruct consumers to provide a rough intensity (e.g. slightly, highly, neither the one nor the other) for these descriptors. However, this would require developing specific data-gathering software. A final software-dependent alternative would be to have bots "talking" with consumers to drive them in their descriptions.

For the study including a temporal component, namely the Free-Comment Attack-Evolution-Finish (FC-AEF) one, the sensory perception was not divided into sensory modalities. The existence of different sensory modalities in the products was rather suggested in the instructions. This decision was taken to avoid the task from being cognitively heavy for the consumers since the sensory perception was already divided into periods. This FC-AEF data gathering procedure appears relevant as it enabled gathering rich temporal characterizations of the products. These characterizations included the information provided by Temporal Dominance of Sensations and Attack-Evolution-Finish (Visalli et al., 2020b) performed on the same products as well as additional information as shown in section B of chapter V and section A of this chapter. However, the assumption that dividing the sensory perception into periods and sensory modalities at the same time would be difficult and cognitively heavy to consumers is questionable. FC-AEF data gathering procedure could be performed with the same rationale as Temporal Dominance of Sensations by Modality (Agudelo et

al., 2015) i.e. one intake by sensory modality for each product. Alternatively, consumers could be instructed to provide for each period one FC description by sensory modality with a single intake of each product. Comparing these different alternatives, including the one presented in this thesis, would be an interesting topic to determine the most performant FC-AEF data gathering procedure.

2. Pretreatment of Free-Comment data

In the proposed pretreatment, two practices enable to standardize the procedure and to reduce the subjective decisions as compared to pretreatment procedures of the previously reported studies (Ares et al., 2010; Hanaei et al., 2015; Lahne et al., 2014; Symoneaux et al., 2012; ten Kleij & Musters, 2003). The first one is the grouping of terms into latent terms containing all their constituting terms displayed. This enables to clarify the groupings of terms that were performed during the pretreatment and to avoid the arbitrary choices of one term to represent several ones. Besides, this enables not to lose a part of the richness of the FC method. However, this proposition can turn out to be a problematic practice for the mapping of the product configuration when too many terms are grouped in a single one. Indeed, the size of labels of the latent descriptors can rapidly increase in such cases and thus obstructing the reading of the map. However, based on the studies of this thesis, this issue appears not to be common, as it never happened. The second practice that enables standardizing the pretreatment procedure is the classification at the grouping of terms step. This reduces the subjectivity of this (almost) necessary step of the pretreatment as groupings are only performed within classes and not between classes. Besides reducing the subjectivity, this practice might prevent performing “wrong” groupings since terms having too different profiles (i.e. in different classes) are unlikely to convey similar descriptive sensory information. Of course, this practice only reduces the subjectivity of this step without entirely removing it since groupings performed within classes remain partly subjective. However, this

part of subjectivity appears not to have a huge impact on the information extracted from the pretreatment (Niedomysl & Malmberg, 2009; Symoneaux et al., 2012).

The semi-automatized procedure proposed in this thesis results from trying to find the optimal tradeoff between the spent time to the pretreatment, and the quality and the quantity of descriptive sensory information extracted by this pretreatment.

The proposed pretreatment is relatively fast thanks to the use of computer software enabling the automation of some steps, and thanks to the classification step shortening the grouping of terms step by making "suggestions" of aggregation. This likely results in a faster procedure than the ones of the previously reported studies where the pretreatment is performed by-hand and without classification of terms (Ares et al., 2010; Hanaei et al., 2015; Lahne et al., 2014; Symoneaux et al., 2012; ten Kleij & Musters, 2003). Given that the code was previously scripted, it took on average two hours and up to four hours to pretreat the datasets of this thesis with the proposed procedure. Of course, this time of pretreatment is relatively long as compared to sensory methods based on a pre-established list of descriptors that do not require such pretreatment of their data. However, this additional time of pretreatment in FC is compensated by the time gained by not having to establish a relevant list of sensory descriptors before the experiment. The pretreatment of FC data is likely less time-consuming than properly establishing a list of sensory descriptors, which makes FC a faster method than list-based ones for not well-known product spaces. Further, the quickness of the pretreatment of FC data can be increased across studies on similar products by creating and enriching sensory lexicons. In the end, when lexicons would be enough enriched, this would result in almost instantaneous FC data pretreatment. Building lexicons may also be the aim of some research projects.

Automatizing some steps of the pretreatment made the task faster, but could possibly result in losing some descriptive sensory information. The first source of

loss of information in the proposed pretreatment is the absence of deep reading of the entire corpus of descriptions, which results in the absence of a systematic correction of spelling and typing errors. This, considered together with the step of hapax removing, can possibly result in removing some occurrences of some final sensory descriptors due to these occurrences not being well formatted. In extreme cases, this can possibly lead to miss some potential sensory descriptors. However, recurrent spelling and typing errors "pass" the step of hapax removing and they are thus corrected at the manual steps of the proposed pretreatment procedure. Thus, only a limited proportion of spelling and typing errors are not considered in the final occurrences of the sensory descriptors. This proportion of spelling and typing errors not considered in the final occurrences could be reduced without adding manual steps by using lexicons and/or string distances (Lu, Lin, Wang, Li, & Wang, 2013; Navarro, 2001). The second source of loss of information in the proposed pretreatment is the two steps of terms removing (hapax and 5% of the panel). These steps likely lead to removing some descriptive sensory information. However, it would be very risky to consider such information in the analysis, as it is not consensual, likely too subtle or it may be noise.

In every study of this thesis, the same procedure of pretreatment of FC data was used. Considering the descriptive sensory information extracted from this procedure and its consistency (no apparent contradictions), it appears that this procedure is relevant. However, the proposed procedure was never compared with others in this thesis. Several steps of the proposed procedure could be performed differently or even removed. For example, the two steps of terms removing could be performed with different thresholds or removed to avoid missing any information. The manual steps could be avoided to render the pretreatment fully automatized. The grouping of terms could be performed with other strategies of classification or even without classification (Ares et al., 2010; Hanaei et al., 2015; Lahne et al., 2014; Symoneaux et al., 2012; ten Kleij & Musters, 2003).

Quantifiers (e.g. *very*, *a little*, etc.) could be taken into account together with negations. This could be performed by considering each pair of quantifier and term as a sensory descriptor on its own (ten Kleij & Musters, 2003). Alternatively, for a given descriptor, the occurrences of this descriptor associated with negative (resp. positive) quantifiers could be considered as half (resp. twice) citations of this descriptor. To summarize, several alternatives to the proposed procedure could be performed to pretreat FC data. Investigating the potential impacts on the extracted descriptive sensory information of these alternatives and comparing them would be an interesting topic to shade more light on the benefits and limitations of each alternative.

Finally, the procedure of pretreatment proposed to be applied for descriptive FC could be easily applied and may be relevant for hedonic-oriented FC or other methods of sensory analysis including free descriptions such as ultra-flash profiling or labeled sorting. It may also be used to help practitioners establishing the lists of list-based methods or to help to extract information from social network text data.

C. Statistical analyses of Free-Comment data

The statistical analyses proposed to analyze Free-Comment (FC) data enhance those of previously reported studies (Ares et al., 2010; Hanaei et al., 2015; Lahne et al., 2014; Symoneaux et al., 2012; ten Kleij & Musters, 2003). After pretreatment, the structure of FC data are identical to that of Check-All-That-Apply (CATA) data, thus the proposed FC analyses can be applied to CATA data as well. They provide a formal and standardized procedure with intuitive visual outputs enabling a fast overview of product discrimination and characterization. In the proposed statistical analysis, the dimensionality and the significance of the dependence are systematically investigated and taken into account in subsequent steps of analysis. The product discrimination is investigated thanks to ellipses and the total bootstrap tests. The significant

associations between products and sensory descriptors are investigated thanks to the tests per cell. Finally, the rationale of the multiple-response chi-square framework completes the previous analyses by rendering them more suited than the usual ones for analyzing FC and CATA data.

The multiple-response chi-square framework models the expected proportions under independence as a function of the estimated probability of a product being evaluated and of the estimated probability of a sensory descriptor being cited within an evaluation. This modeling does not explicitly take into account the individual (i.e. consumer) effects nor the joint probabilities of citations between sensory descriptors. In the proposed analyses, the individual effects and the joint probabilities of citations between sensory descriptors are taken into account implicitly thanks to the Monte-Carlo and bootstrap procedures. Explicitly taking into account the individual effects and/or the joint probabilities of citations between sensory descriptors could be an improvement of the multiple-response chi-square framework when used for analyzing sensory data. However, explicitly taking them into account would come with the limitation of increasing the number of parameters estimated while the amount of data available for these estimations would not remarkably increase due to practical limitations. This could result in unstable estimations leading to unstable conclusions. Nevertheless, it remains that investigating a more complex model and comparing it to the proposed analyses would be an interesting topic for future research. This could help to determine to which extent explicitly taking into account the individual effects and/or the joint probabilities of citations between sensory descriptors is, or not, beneficial over implicitly taking them into account.

In the proposed analyses, the product configuration is depicted by multiple-response correspondence analysis. This strategy can be summarized as follows: compute a distance at the panel level for each pair of products and find the ranked orthogonal axes that best retrieve these distances. This fits into the following

rationale: it exists “true” sensory distances between products and the distances computed at the panel level is a good estimator of these “true” sensory distances. Another common rationale in sensory analysis is the following: it exists actual latent sensory dimensions in the product space and each consumer perceives or not each of these dimensions with a more or less high degree of importance. The strategy of analysis corresponding to this rationale is to find the latent sensory dimensions shared by a maximum of consumers and considering them as a good approximation of the actual latent sensory dimensions of the product space. To apply this strategy to FC data, different methods could be considered with adaptations such as Generalized Procrustes Analysis (GPA) (Gower, 1975), Multiple Factor Analysis (MFA) (Escofier & Pagès, 1994), Structuration des Tableaux A Trois Indices de la Statistique (STATIS) (Lavit, Escoufier, Sabatier, & Traissac, 1994), or Common Component and Specific Weight Analysis (CCSWA) (Qannari, Wakeling, Courcoux, & MacFie, 2000). To the best of our knowledge, no application of analyses fitting into this second rationale to FC data has been reported. Comparing these two rationales by investigating their respective benefits and limitations for analyzing FC data, and more generally descriptive sensory data, would be an interesting topic for future research.

Finally, all analyses proposed in this thesis to analyze FC data are multidimensional i.e. they investigate the differences between the products based on the entire set of sensory descriptors at the same time. Alternatively, unidimensional analyses, such as Cochran’s Q test (Cochran, 1950) or generalized linear model could be performed to investigate the differences between the products by descriptor. Unidimensional analyses have been set aside in this thesis because sensory perception is essentially a multidimensional phenomenon in which each dimension modifies the perception of the others. For example, when dealing with basic tastes, the perception of sweetness is often counterbalanced by this of sourness. It is thus very likely that if differences are found between the

products concerning sweetness, differences between products regarding sourness will be found also. These kinds of interconnection likely occur between several descriptors and probably with more than two descriptors at a time and thus explains why multidimensional analyses were proposed with unidimensional ones set aside.

D. Performances of Free-Comment as compared to Check-All-That-Apply

1. Discrimination and characterization of the products

The study of this thesis that compared Free-Comment (FC) and Check-All-That-Apply (CATA) in terms of discrimination and characterization of the products showed that FC performed well and is even better than CATA regarding this feature in some situations. Indeed, FC overall provided better discrimination as well as richer and more precise characterization of the products than CATA.

For this comparison, the pre-established list of descriptors for the CATA method was based on the expertise of wine professionals. Consequently, some descriptors of the list might have been perceived as somewhat “technical” by the consumers. This could potentially have impeded the ability of CATA to discriminate and precisely characterize the products. However, several descriptors of the CATA list were mentioned in the FC descriptions (e.g. opaque, black, bright, animal, balanced, astringent, etc.), suggesting that they likely made sense for the consumers.

The comparison of FC and CATA regarding their ability to discriminate and characterize the products was performed by sensory modalities (visual, olfactory and gustatory). This division may have favored FC as consumers were driven into characterizing every sensory modality of the products. Without this driving, consumers may have mentioned only one or two of the sensory modalities in their FC descriptions, resulting in less accurate discrimination and

characterization of the products. On the other hand, CATA without division of the sensory perception would likely have performed similarly because the list of descriptors would have driven consumers to describe every sensory modality. However, without division of the sensory perception, the list of descriptors in CATA would have looked oversized (37 descriptors), and thus demotivated the consumers. Even if FC could potentially perform less well than CATA without division of the sensory perception into sensory modalities, which is still an open question, it remains that FC appears to perform at least as well if not better when this division is performed. This, considered together with the fact that dividing the sensory perception is not costly from any point of view, suggests that replacing CATA with FC in practical applications might be relevant. However, this remains to be confirmed by future studies.

Speaking of other studies, the recent study (not yet published at the time this manuscript is written) already mentioned in section A of this chapter added information on this topic. As a reminder, this study involved 98 consumers that performed a FC on 4 dark chocolates and the resulting *a posteriori* list of sensory descriptors was used into a CATA performed by these same 98 consumers on the same 4 dark chocolates two weeks later. FC and CATA were both in blind testing conditions. The chocolates were characterized according to their texture in mouth and their flavor in mouth. In this latter study, the differences in performance between FC and CATA were less clear. Both methods overall provided similar insights on the products but with a lesser agreement and some contradictions regarding the texture modality. CATA provided a more important level of dependence between products and sensory descriptors as well as more product by descriptor significant associations. This, considered together with the acquiescence bias highlighted in section A of this chapter and the contradictions with FC despite the use of the same list of descriptors, reinforces the hypothesis that CATA might be more “powerful” (easily detect present sensory

characteristics) but also more "risky" (easily detect questionable sensory characteristics) than FC. Verifying this hypothesis based on products for which sensory characteristics are controlled would definitely be of great interest. Despite CATA could perform better than FC when its list is based on previous FC descriptions, the FC descriptions already provide most of the information in a less time-consuming, less expensive and less biased fashion. This, considered together with other benefits of FC, again suggests that replacing CATA with FC in practical applications might be relevant.

2. Stability of the provided descriptive sensory information

The study of this thesis that compared FC and CATA in terms of stability of the provided descriptive sensory information showed that FC provides relatively stable information, at least as stable as CATA and even slightly more in some cases.

To compare FC and CATA in terms of stability of the provided descriptive sensory information, three criteria were investigated: stability of the product configuration, stability of the joint descriptor by product configuration, and stability of the joint descriptor by product significant associations. While the stability of the product configuration is systematically investigated in studies of stability (Ares, Bruzzone, et al., 2014; Ares, Tárrega, Izquierdo, & Jaeger, 2014; Blancher, Clavier, Egoroff, Duineveld, & Parcon, 2012; Cadena et al., 2014; Vidal, Tárrega, Antúnez, Ares, & Jaeger, 2015), it was the first time, to the best of our knowledge, that the stability of the sensory interpretation was investigated in a sensory study with consumers. However, the stability of the sensory interpretation had been studied before in the context of sensory profiling (Heymann et al., 2012; Peltier, Mammasse, Visalli, Cordelle, & Schlich, 2018). In the context of consumer methods of sensory analysis, some authors investigated the stability of the descriptor (Ares, Bruzzone, et al., 2014; Ares,

Tárrega, et al., 2014; Vidal et al., 2015) but did not put it in regards to the sensory interpretation of product differences.

In this study, whatever the criterion considered, the information provided by the actual panel was considered as a benchmark to which was compared the information provided by the bootstrapped virtual panels. Despite this is the usual way of investigating stability (Ares, Bruzzone, et al., 2014; Ares, Tárrega, et al., 2014; Blancher et al., 2012; Cadena et al., 2014; Heymann et al., 2012; Mammasse & Schlich, 2014; Peltier et al., 2018; Vidal et al., 2015), alternatives could be thought of. A first alternative would require conducting the study with a relatively large panel (say, 120 consumers). Based on this large panel, the information provided by bootstrapped virtual panels (say, of 60 consumers) with no intersection of consumers could be compared instead of considering the actual panel as a benchmark. This alternative would show the benefit of ensuring no intersection between the consumers of the two compared panels and thus providing more realistic comparisons. However, a limitation of this alternative is that it requires a large actual panel, which can turn out time-consuming and expensive from a practical point of view. A second alternative would be no longer to compare the actual and virtual panels two by two but rather investigate the overall agreement of several panels similarly to what Structuration des Tableaux A Trois Indices de la Statistique (STATIS) (Lavit et al., 1994) performs at the individual level. The measure of similarity would no longer be the correlation coefficients but the first eigenvalue (eventually reduced to a percentage) of the matrix containing pairwise (generalized) correlation coefficients between panels. This alternative shows the benefit of approximating the overall level of stability whatever the chosen subsample of consumers. However, repeating this procedure a large number of times is computationally expensive and thus time-consuming. Indeed, this requires building several virtual panels, then investigating their overall agreement, and finally repeating these two steps a large number of times.

The two previously mentioned alternatives could even be combined into a third alternative. To summarize, several alternatives to the usual reported one could be applied in studies of stability. Comparing these alternatives by investigating their respective benefits and limitations would be an interesting topic to determine the most suited approach for studying stability.

In this study of stability, the limitations mentioned in the previous section regarding the wine datasets also apply. Regarding the chocolate datasets, the CATA was performed by consumers that might be more knowledgeable about chocolate than those who performed the FC. Thus, the CATA descriptions might have been more consensual, which might have resulted in higher stability of the outputs. Therefore, the stability of CATA outputs might have been overestimated for the chocolate datasets. Anyway, the conclusions from this study remain to be confirmed by future studies.

E. Extensions of Free-Comment

1. Temporal sensory analysis

The application of FC-AEF on dark chocolates proposed in this thesis showed that FC-AEF performs well. Indeed, it was able to provide temporal discrimination and characterization of the products based on the descriptors from Visalli et al. (2020b) retrieved in the FC descriptions as well as additional descriptors as mentioned in section A of this chapter. These additional descriptors exhibit the benefits of FC over pre-established lists of sensory descriptors in the context of temporal sensory analysis.

By being based on AEF, FC-AEF shares some of its features and thus the benefits and limitations associated with these features (Visalli et al., 2020b). The first feature is the discretization of the time *a priori* into three periods. This comes in opposition to the continuous-time of the most common temporal methods of sensory analysis conducted with consumers, namely Temporal Dominance of

Sensations (TDS) (Pineau et al., 2003; Pineau et al., 2009) and Temporal-Check-All-That-Apply (TCATA) (Castura et al., 2016). Discretizing the time results in gathering fewer data, which may result in losing information in some cases as the number of periods considered is necessarily limited and no durations are recorded. However, discretizing the time also shows the benefit of suppressing any individual differences in terms of response delays, mind processing of the sensory perception, and duration of the sensory perception, which can noise the information in continuous-time methods. The relative amount of information lost as compared to the noise left aside by discretizing the time remains an open question deserving of being investigated in future studies.

The second feature FC-AEF inherited from AEF is its retrospective reporting of the temporal sensory perception as opposed to TDS and TCATA where the reporting is expected to be concurrent to the perception. This retrospective aspect is necessary to perform FC-based temporal sensory analysis because typing concurrently to the perception is not realistic, at least for products eliciting relatively short perceptions and/or with relatively fast kinetics. However, this retrospective reporting necessarily involves the memory of the consumers, which can be seen as a non-desired additional cognitive effort. Involving memory may result in a less-instinctive and more mind-processed reported temporal sensory perception. This shows the limitation of gathering information that may be further apart from “reality” but also shows the benefit of gathering what consumers remember from their experience with the product. This memory might be more relevant in some situations as it represents what consumers have in their mind when comes to decide which product to buy or discuss it with other people. The retrospective aspect also shows a practical benefit: the task does not require to be extensively briefed to consumers as opposed to TDS and TCATA, which require to be well explained to be understood (Visalli et al., 2020b). Indeed, first evaluating and then reporting their perception is what panelized consumers are

used to performing whereas reporting while evaluating is more unusual and requires specific actions from them.

In the application of FC-AEF proposed in this thesis and similarly to the non-temporal applications of FC, consumers were not restricted in terms of the number of terms they could provide in their FC descriptions. This application of FC-AEF thus positions it close to the rationale of "applicability" from TCATA, where several descriptors can be mentioned at each time. An alternative version of FC-AEF closer to the rationale of "dominance" from TDS could be used as well. In this alternative, consumers would be restricted to a single term or expression for each of the three periods. This restriction would show the benefit of reducing the memory effort required from consumers as well as catching the most salient information and standardizing it (same number of descriptors for each triple of consumer, product and period) but at the cost of a possible loss of information.

FC-AEF provides for each period a proportion of citations of each descriptor for each product. Two points of view have been proposed to analyze these data: products by period and periods by product. The first proposition intends to answer the question: "Do the products differ at a given period?". The second proposition intends to answer the question: "Do the periods of a given product differ?". These two analysis strategies show the benefit of providing a direct answer to the two questions. However, the three-way nature of the data is not taken into account. Several methods could be applied to acknowledge the three-way nature of the data, such as parallel factor analysis (PARAFAC) (Harshman & Lundy, 1994), or GPA, MFA, STATIS, CCSWA, methods already mentioned in this manuscript. These methods could be useful to determine and investigate shared or not sensory dimensions between the periods. Figure 3 shows an example of the application of PARAFAC on FC-AEF data.

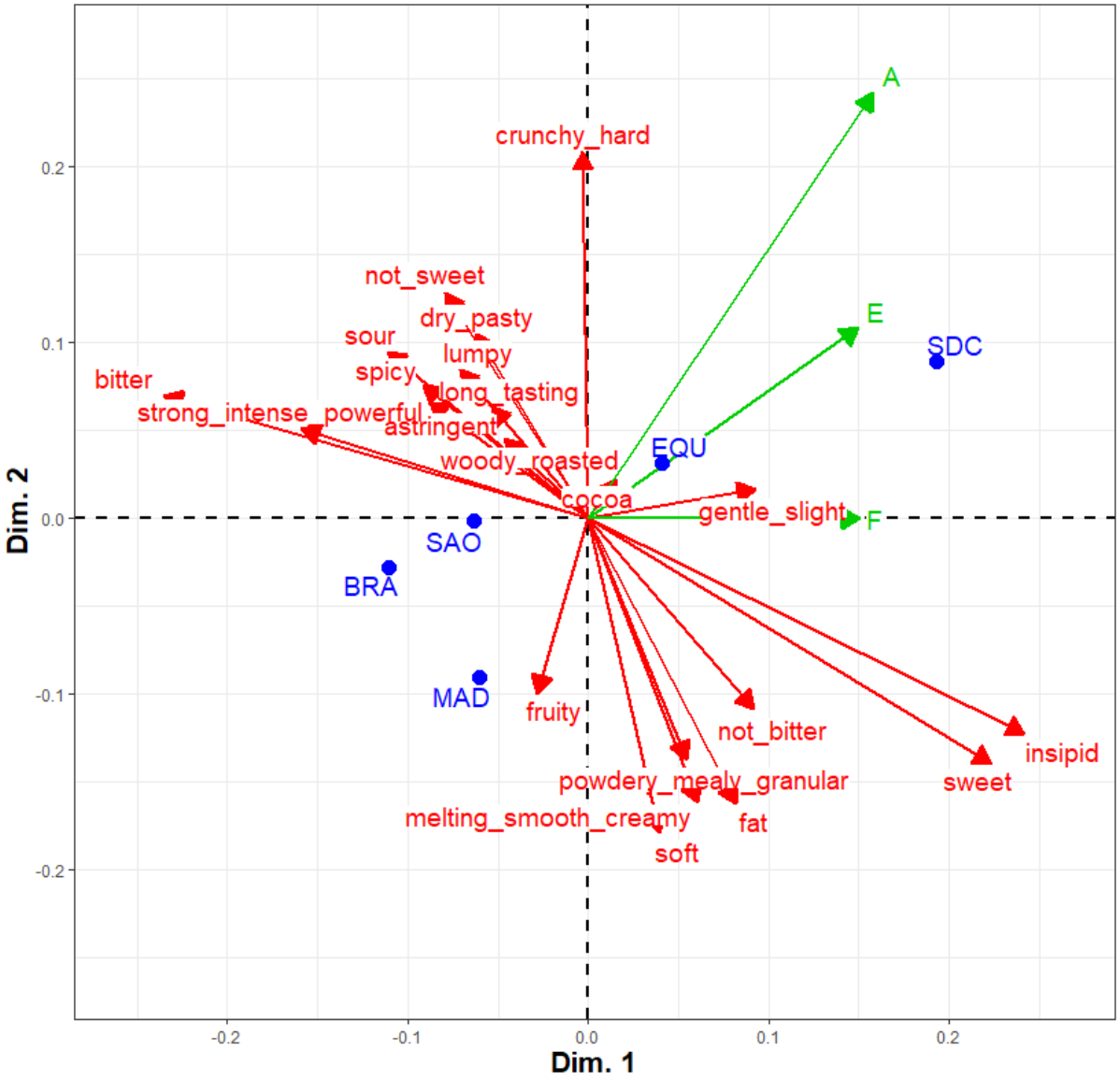


Figure 3: Example of PARAFAC applied on FC-AEF data

Input matrices were the multiple-response chi-square standardized residual with Mode A being the products, Mode B being the descriptors, and Mode C being the periods. Loadings of mode C (periods) were constrained to be positive and the model was fitted considering two components. In this example, the first axis is mainly a gradient of flavor strength between the products associated with aromas and basic tastes. The second axis is mainly a gradient of texture. Interestingly, the loadings of the periods are very similar on the first axis suggesting that the gradient of flavor strength is constant across periods. On the second axis, the loadings of the periods are decreasing from A to F highlighting logically that the

texture dimension becomes less perceptible along the intake. This makes sense since products are destructed during mastication, which progressively erases their differences in terms of texture.

Another strategy to acknowledge the three-way nature of the data would be to derive sensory trajectories based on two-way methods such as Principal Component Analysis (Castura et al., 2016; Lenfant, Loret, Pineau, Hartmann, & Martin, 2009; Visalli et al., 2020b). Sensory trajectories could be useful to provide an overview of between and within products temporal sensory kinetics at the same time. This might be relevant to directly identifying products having or not similar temporal kinetics and sensory characteristics driving them. Figure 4 shows an example of sensory trajectories derived from FC-AEF data and using multiple-response Correspondence Analysis.

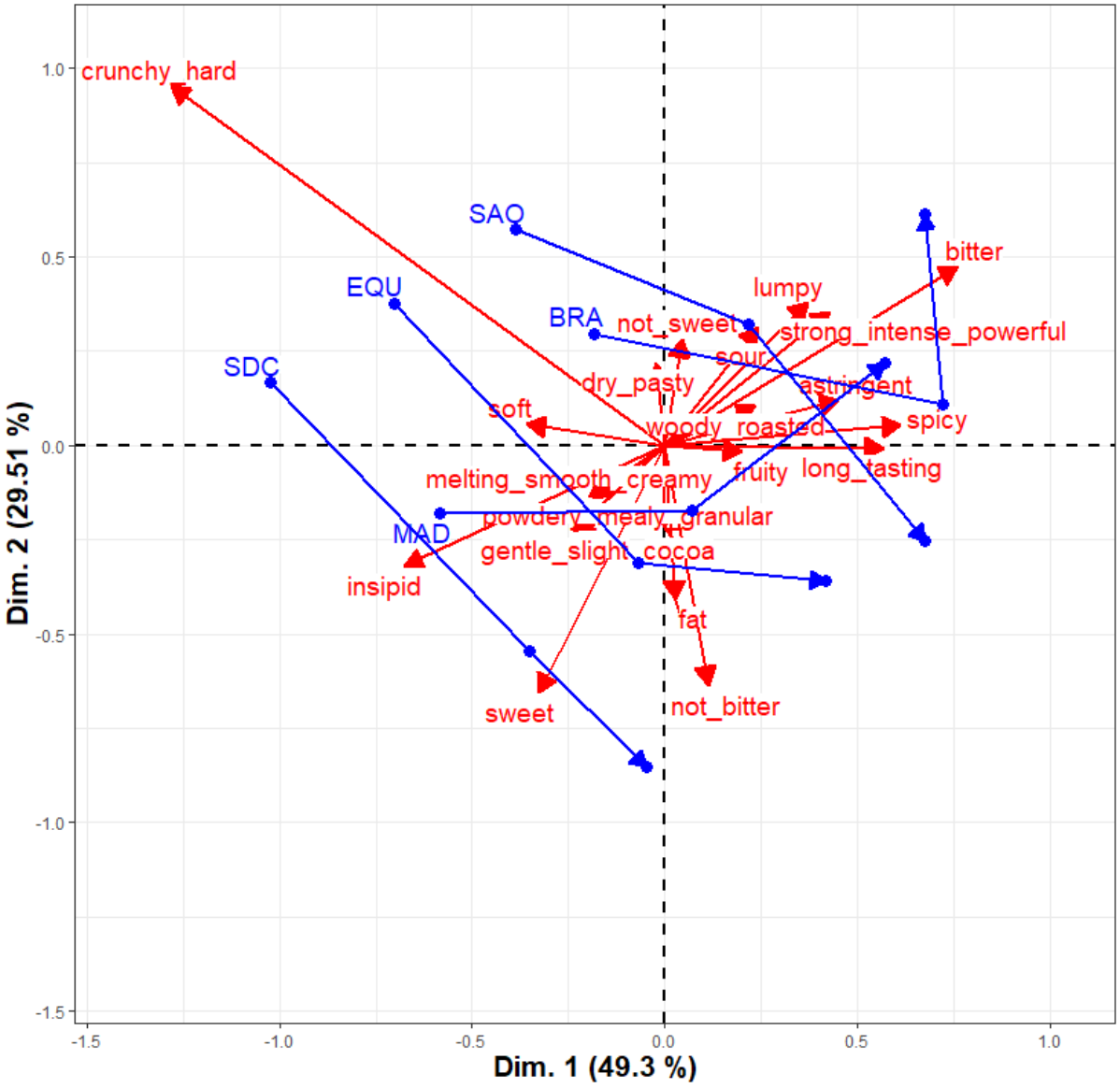


Figure 4: Example of sensory trajectories derived from FC-AEF data

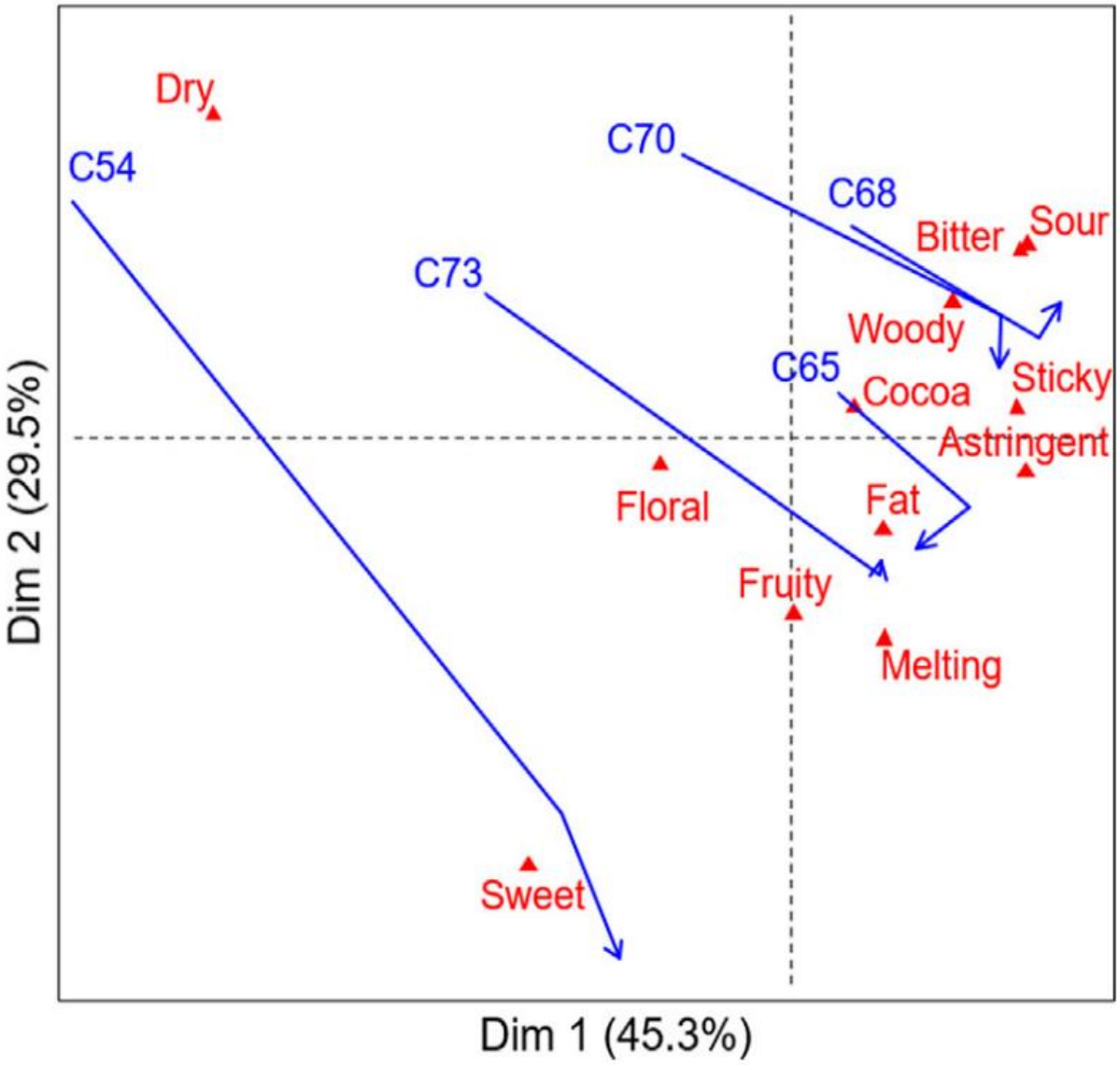


Figure 5: Sensory trajectories derived from AEF data retrieved from Visalli et al. (2020b)

In this example in Figure 4, each product shows a specific location and trajectory. Interestingly, the three main sensory poles of sensory trajectories derived from AEF in Visalli et al. (2020b) and depicted in Figure 5, namely the "*sweet"* one, the "*dry"* one and the "*bitter, sour, woody"* one are rediscovered by the FC-AEF. However, two important differences of sensory trajectories derived from FC-AEF as compared to those derived from AEF are to be noticed. First, the characterizations of the poles are richer and/or more precise: the "*dry*" one is rather a *crunchy_hard* one (c.f. section A of this chapter), the "*sweet*" one is rather a *sweet, insipid, not_bitter* one and the "*bitter, sour, woody"* one is

associated with complementing information such as *long_tasting*, *strong_intense_powerful*, *spicy*, etc. Second, trajectories derived from FC-AEF are less flat, less parallel and more complex than those derived from AEF. These assessments confirm that using FC rather than pre-established lists of sensory descriptors in the context of temporal sensory analysis might be beneficial by providing more precise and accurate descriptive characterizations.

Investigating the benefits and the limitations of the different strategies of analysis applicable to FC-AEF data would be an interesting topic to determine which strategy of analysis is advisable depending on the aims of the study. Finally, FC-AEF data can be analyzed as static FC data by aggregating periods by product. This strategy could be undertaken when the temporality is not well marked.

2. Drivers of liking identification and ideal product characterization

To extend FC to drivers of liking identification and ideal product characterization, the Ideal-Free-Comment (IFC) method has been introduced and proposed to be paired with liking scoring. In this method, three types of data are gathered from the consumers: FC descriptions of the actual products under interest, liking scores of these products and FC descriptions of the virtual ideal product. Three strategies of analysis were proposed with these data. The first strategy consists in regressing liking scores of the actual products on the corresponding FC descriptions to identify drivers of liking. The second strategy consists in estimating the proportion of citations of each sensory descriptor for the ideal product and testing them against the corresponding proportions for the pool of actual products. This enables to characterize the ideal product and its differences from the actual products. The third strategy consists in projecting the ideal product and the mean liking scores of the actual products in the sensory space depicted derived from the characterizations of the actual products. This enables to locate the ideal product relatively the actual products and their liking scores. The application of IFC paired with liking scoring on cooked hams

proposed in this thesis showed this method able to fulfill the aims for which it was designed. Indeed, the identified positive and negative drivers of liking made sense and the ideal product characterization was consistent with them and with liking scores of the actual products.

Drivers of liking identification and ideal product characterization have been proposed as complementary tools to provide insights on consumers' preferences based on FC data. Ideal product characterization benefits from investigating a larger sensory space than the one defined by the actual products of the study, which is relevant in some practical applications (Worch, Crine, Gruel, & Lê, 2014), but it suffers from cognitive and attitudinal biases in some others (Li, Hayes, & Ziegler, 2015). Drivers of liking are identified *a posteriori* without consumers being aware of this procedure, which shows the benefit of being implicit and thus not cognitively biased. However, drivers of liking depend on the actual products, which could result in some loss and/or some misleading information if too many sensory characteristics are confused and/or not well represented by the actual products.

Drivers of liking have been proposed to be identified based on a mixed linear model and using the resulting loadings with their confidence intervals. Alternatively, "*penalty-lift analysis*" (Meyners, Castura, & Carr, 2013; Williams, Carr, & Popper, 2011) may have been performed. However, *penalty-lift analysis* suffers from not accounting for eventual correlations between descriptors (Meyners et al., 2013) while the mixed linear model does to some extent. Further, using a mixed linear model enables to account for the consumer and product effects which *penalty-lift-analysis* does not. In the application of this thesis, the product effect was considered as fixed in the model. This decision was taken considering that the products of the study were specifically selected to span the variability of several fixed effects (salt content, fat content, absence or presence of several specific labels) of the market. In applications where the products of the

study would be selected more or less arbitrarily among the product category, considering the product effect as random in the mixed linear model might be more relevant.

As the ideal product characterization can be cognitively biased in some situations, checking the consistency of these data is important (Worch, Lê, Punter, & Pagès, 2012a, 2012b). In the proposed application, the consistency of the ideal product data was investigated two ways. First, the ideal product characterization was compared with the drivers of liking. Second, the location of the ideal product in the sensory space was compared to the direction of increasing liking in this space. Alternatively or complimentarily, the predicted liking scores of the ideal product from a model fitted on the liking scores of the actual products and their sensory descriptions may have been performed (Worch et al., 2012a, 2012b). The results of this analysis are depicted on Figure 6 to complement the results from the study.

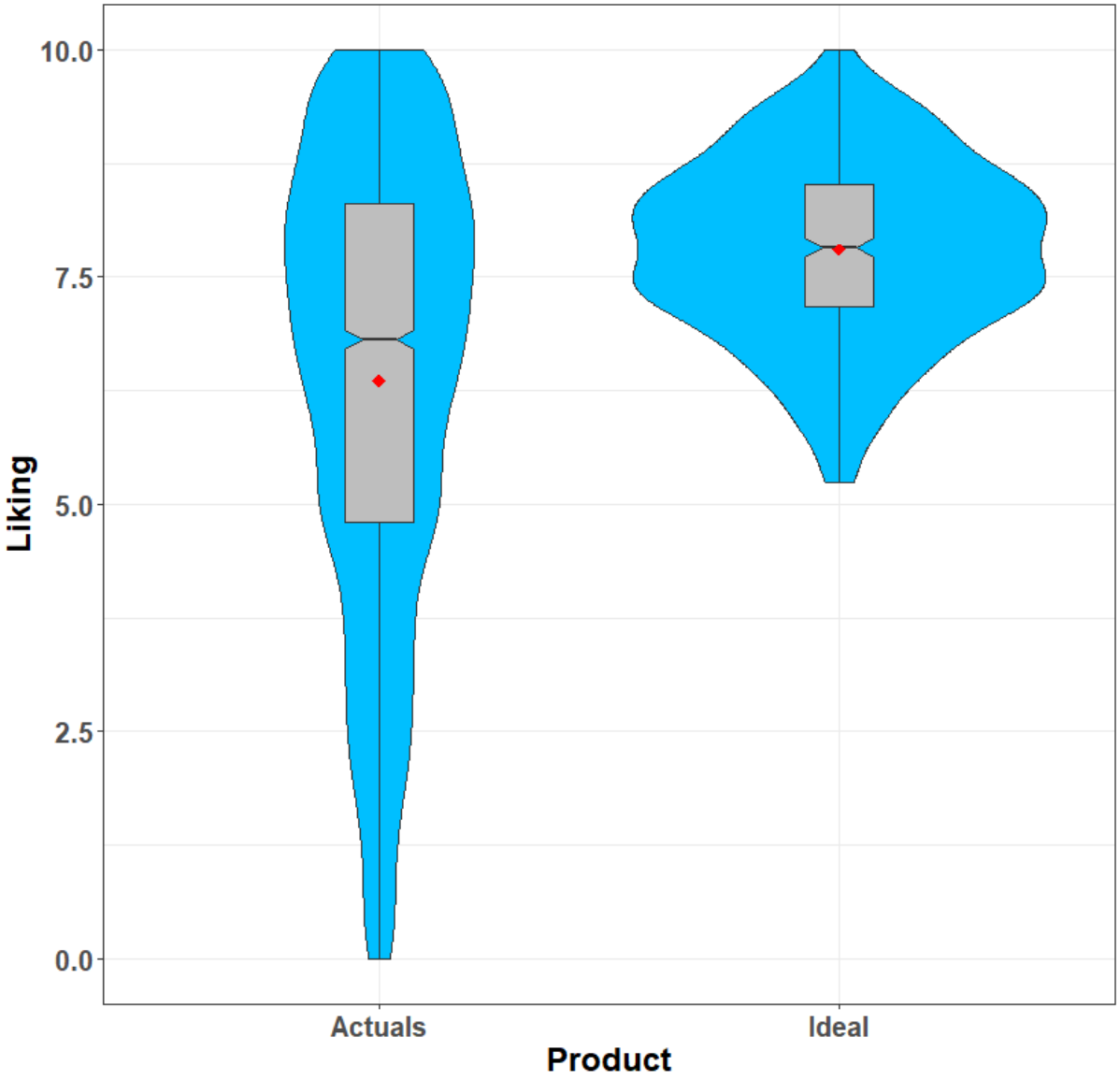


Figure 6: Distribution of liking scores for the pool of actual products (measured) and for the ideal product (predicted) from the ham study of chapter V. The red mark indicates the mean of the corresponding distribution

The liking scores of the actual products were regressed against the consumer factor and the descriptor factors using a mixed linear model fitted on the 2758 evaluations. Each descriptor factor had two levels: absence or presence, the absence level being the reference one. The descriptor factors were considered as fixed while the consumer factor was considered as random. The resulting model was applied to the ideal product data to obtain potential liking scores for this product. This additional analysis confirms that the ideal product data were

consistent in this study as the predicted liking scores for the ideal product had a higher mean and median than the liking score for the actual products.

Despite the possibility of checking the consistency of the ideal product data and their recurrent consistency (Ares, Varela, et al., 2011; Brard & Lê, 2016; Worch et al., 2014; Worch et al., 2012a, 2012b), ideal product characterization is sometimes criticized. Whatever ones' opinion on the ideal product characterization, instructing consumers to describe their ideal product after having evaluated every product of the study does not cost anything, it has no impact on the gathered data of the actual products and it often provides relevant information. Based on these assessments, why not instructing consumers to describe their ideal product systematically?

In the proposed application, the analysis was performed at the product space level and was not specific to any product. However, in practical applications, it is common to need by-product information for the hedonic optimization of some products in particular. A first alternative to obtain such information could be to use the valence of the sensory descriptors together with the degree of association of the product under interest with these descriptors. This information can for example coming from the same model used to identify drivers of liking for the valence, and from the multiple-response chi-square standardized residual matrix for the degree of association. If the product under interest is positively (resp. negatively) associated with some descriptors having a negative (resp. positive) valence, then this gives directions for hedonic optimization of this product. A second alternative to obtain by-product information could be to compare the sensory characterization of the product under interest to the ideal product characterization in terms of citation proportions of the sensory descriptors (Meyners et al., 2013). Of course, as in the proposed application, the two previous alternatives could be performed conjointly as they might be complementary.

Another strategy of analysis to exploit IFC and hedonic data could be performed. This strategy would be to compute the deviation of each product from the ideal on each descriptor and then to consider the resulting information as Just-About-Right (JAR) data (Meyners et al., 2013; Worch, Dooley, Meullenet, & Punter, 2010). Then the analyses usually applied on JAR data, notably "*penalty analysis*", could be applied equally to these JAR encoded data. *Penalty analysis* can be seen as standing at the frontier between drivers of liking identification and ideal product characterization and thus aggregating their respective benefits, but unfortunately, their respective limitations. Investigating the practical benefits and limitations of *penalty analysis* over the analyses proposed in the application of this thesis and these mentioned above could be an interesting topic for future research. Further, investigating whether the JAR encoded IFC data and the Free JAR data (Luc et al., 2020) lead to a similar ideal product and provide similar directions of improvement could be an interesting complement to this research topic. This could help to determine which methods are advisable over the other ones to gather relevant information on drivers of liking, ideal product and direction of improvement thanks to FC.

The approaches proposed in the application of this thesis as well as those mentioned above to understand consumers' appreciation thanks to FC might deserve to be applied and compared based on a more balanced design than the one of the ham study. In such a protocol, every consumer involved in the study would evaluate every product under interest.

If one is skeptical regarding all the approaches mentioned above but still interested in understanding consumers' appreciation thanks to FC, it must be quoted that "classical" preference mapping techniques (Carroll, 1972; Danzart, 2009; Greenhoff & MacFie, 1994; McEwan, 1996; Schlich & McEwan, 1992) could be performed based on FC sensory data and hedonic data. Further, hedonic-oriented FC where consumers are instructed to categorize their descriptions into

a "*like*" category and a "*dislike*" one (Lahne et al., 2014; Symoneaux et al., 2012) might be considered as another alternative. These two alternatives could be interesting to be compared to ideal-related methods based on FC.

Finally, all the approaches discussed above are based on an explicit measurement of the hedonic appreciation, i.e. liking scores provided by the consumers. Alternatively, these approaches could likely be applied based on an implicit measurement of the hedonic appreciation, coming from the sentiment analysis of the descriptions (Luc et al., 2020; Visalli, Mahieu, Thomas, & Schlich, 2020a) for example. Comparing the results obtained either based on an explicit measurement or an implicit measurement of the hedonic appreciation might be interesting to determine the type of measurement that is the most relevant and representative to measure hedonic appreciation.

Chapter VII: Conclusion

This thesis aimed to put Free-Comment (FC) in the spotlight for sensory analysis with consumers as it shows several benefits. The most notable benefits of FC are to avoid inherent limitations to pre-established lists of sensory descriptors, to be easy to set up and to be easy to perform for consumers. Its natural and spontaneous aspect renders FC flexible and self-explicit for consumers, which makes it a relevant method for less controlled testing conditions such as home-used tests. Further, FC has the additional benefit of having its data aggregable across different studies. Particular attention was given to practicality in the propositions made along this thesis.

FC data gathering has been proposed to be performed by sensory modality, i.e. consumers provide one description for each investigated sensory modality in the products under interest. This approach enables exploiting as much as possible the capacities of FC resulting in a rich characterization of the products regarding their investigated sensory modalities.

For the pretreatment of FC data, a semi-automatized procedure has been proposed. This procedure enables to standardize and simplify as much as possible this necessary step of FC data analysis. It aims to offer a good balance between keeping analysis time reasonable and minimizing the loss of information.

For the statistical analysis of FC data, operating in the significant subspace of product by sensory descriptor dependences is proposed together with the multiple-response chi-square framework that better takes into account the structure of the pre-treated data than the usual chi-square framework. The resulting analyses enable deeper exploitation of FC data within a more suited framework than the usual ones together with easy to interpret outputs. These analyses have been implemented into the MultiResponseR R-package, which is presented in Appendix and freely available at: https://github.com/MahieuB/MultiResponseR. Remember that since the

pretreated FC data have the same structure as the Check-All-That-Apply (CATA) data, the proposed analyses are also relevant for analyzing CATA data.

FC has been compared to CATA on two performance criteria: the ability to discriminate and characterize the products and the ability to provide stable descriptive sensory information. Regarding both criteria, FC turned out to perform at least as well as CATA, if not better. Indeed, it provided better discrimination and richer characterization with slightly higher stability. This suggests no loss of performance goes along with the benefits of FC.

An extension of FC to temporal sensory analysis, called Free-Comment Attack-Evolution-Finish (FC-AEF), has been introduced. FC-AEF enables catching the temporal kinetic of the sensory perception paired with the benefits of FC. An application of FC-AEF demonstrated its ability to provide temporal sensory discrimination and characterization of the products.

An extension of FC to drivers of liking identification and ideal product characterization, called Ideal-Free-Comment (IFC) paired with liking scoring, has been introduced. This method enables to investigate consumers' hedonic appreciation based on FC data and thus without the limitations inherent to a pre-established list of sensory descriptors and by maximizing the chances of not missing key information. An application of IFC paired with liking scoring demonstrated its ability to identify relevant drivers of liking congruent with ideal product characteristics, these two tools being complementary.

Overall, this work demonstrated the potency and the versatility of the Free-Comment method. It opens new perspectives for sensory analysis with consumers and it should promote a larger use of Free-Comment in that field.

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.

Agudelo, A., Varela, P., & Fiszman, S. (2015). Methods for a deeper understanding of the sensory perception of fruit fillings. *Food Hydrocolloids, 46*, 160-171.

Albert, A., Salvador, A., Schlich, P., & Fiszman, S. (2012). Comparison between temporal dominance of sensations (TDS) and key-attribute sensory profiling for evaluating solid food with contrasting textural layers: Fish sticks. *Food Quality and Preference, 24*(1), 111-118.

Ares, G., Bruzzone, F., & Giménez, A. N. A. (2011). Is a consumer panel able to reliably evaluate the texture of dairy desserts using unstructured intensity scales? Evaluation of global and individual performance. *Journal of Sensory Studies, 26*(5), 363-370.

Ares, G., Bruzzone, F., Vidal, L., Cadena, R. S., Giménez, A., Pineau, B., et al. (2014). Evaluation of a rating-based variant of check-all-that-apply questions: Rate-all-that-apply (RATA). *Food Quality and Preference, 36*, 87-95.

Ares, G., Castura, J. C., Antúnez, L., Vidal, L., Giménez, A., Coste, B., et al. (2016). Comparison of two TCATA variants for dynamic sensory characterization of food products. *Food Quality and Preference, 54*, 160-172.

Ares, G., Dauber, C., Fernandez, E., Gimenez, A., & Varela, P. (2014). Penalty analysis based on CATA questions to identify drivers of liking and directions for product reformulation. *Food Quality and Preference, 32*, 65-76.

Ares, G., de Andrade, J. C., Antúnez, L., Alcaire, F., Swaney-Stueve, M., Gordon, S., et al. (2017). Hedonic product optimisation: CATA questions as alternatives to JAR scales. *Food Quality and Preference, 55*, 67-78.

Ares, G., Giménez, A., Barreiro, C., & Gámbaro, A. (2010). Use of an open-ended question to identify drivers of liking of milk desserts. Comparison with preference mapping techniques. *Food Quality and Preference, 21*(3), 286-294.

Ares, G., Jaeger, S. R., Bava, C. M., Chheang, S. L., Jin, D., Gimenez, A., et al. (2013). CATA questions for sensory product characterization: Raising awareness of biases. *Food Quality and Preference, 30*(2), 114-127.

Ares, G., Tárrega, A., Izquierdo, L., & Jaeger, S. R. (2014). Investigation of the number of consumers necessary to obtain stable sample and descriptor

configurations from check-all-that-apply (CATA) questions. *Food Quality and Preference, 31*, 135-141.

Ares, G., Varela, P., Rado, G., & Giménez, A. (2011). Identifying ideal products using three different consumer profiling methodologies. Comparison with external preference mapping. *Food Quality and Preference, 22*(6), 581-591.

Ayidiya, S. A., & McClendon, M. J. (1990). Response effects in mail surveys. *Public Opinion Quarterly, 54*(2), 229-247.

Becker, S. L. (1954). Why an Order Effect. *The Public Opinion Quarterly, 18*(3), 271-278.

Bécue-Bertaut, M., & Lê, S. (2011). Analysis of multilingual labeled sorting tasks: application to a cross-cultural study in wine industry. *Journal of Sensory Studies, 26*(5), 299-310.

Benzécri, J.-P. (1973). *Analyse des données. Analyse des correspondances (Vol. 2)* Paris: Dunod.

Blancher, G., Clavier, B., Egoroff, C., Duineveld, K., & Parcon, J. (2012). A method to investigate the stability of a sorting map. *Food Quality and Preference, 23*(1), 36-43.

Bradburn, N., & Sudman, S. (1979). *Improving Interview Method and Questionnaire Design*.

Brand, J., Valentin, D., Kidd, M., Vivier, M. A., Næs, T., & Nieuwoudt, H. H. (2020). Comparison of pivot profile© to frequency of attribute citation: Analysis of complex products with trained assessors. *Food Quality and Preference, 84*, 103921.

Brandt, M. A., Skinner, E. Z., & Coleman, J. A. (1963). Texture Profile Method. *Journal of Food Science, 28*(4), 404-409.

Brard, M., & Lê, S. (2016). The Ideal Pair Method, an Alternative to the Ideal Profile Method Based on Pairwise Comparisons: Application to a Panel of Children. *Journal of Sensory Studies, 31*(4), 306-313.

Bruzzone, F., Vidal, L., Antúnez, L., Giménez, A., Deliza, R., & Ares, G. (2015). Comparison of intensity scales and CATA questions in new product development: Sensory characterisation and directions for product reformulation of milk desserts. *Food Quality and Preference, 44*, 183-193.

Cadena, R. S., Caimi, D., Jaunarena, I., Lorenzo, I., Vidal, L., Ares, G., et al. (2014). Comparison of rapid sensory characterization methodologies for the development of functional yogurts. *Food Research International, 64*, 446-455.

Cairncross, S. E., & Sjostrom, L. B. (1950). Flavor profiles: A new approach to flavor problems. *Food Technology, 4*, 308-311.

Callegaro, M., Murakami, M. H., Tepman, Z., & Henderson, V. (2015). Yes-no answers versus check-all in self-administered modes. *International Journal of Market Research, 57*, 203-223.

Camiz, S., & Gomes, G. C. (2013). Joint Correspondence Analysis Versus Multiple Correspondence Analysis: A Solution to an Undetected Problem. In, *Classification and Data Mining*.

Campbell, D. T., & Mohr, P. J. (1950). The effect of ordinal position upon responses to items in a check list. *Journal of Applied Psychology, 34*(1), 62-67.

Campo, E., Ballester, J., Langlois, J., Dacremont, C., & Valentin, D. (2010). Comparison of conventional descriptive analysis and a citation frequency-based descriptive method for odor profiling: An application to Burgundy Pinot noir wines. *Food Quality and Preference, 21*(1), 44-55.

Carroll, J. D. (1972). Individual Differences and Multidimensional Scaling. In, *R. N. Shepard, A. K. Romney, & Si Nerloves (Eds.), Multidimensional scaling: Theory and applications in the behavioral sciences*. New York: Academic Press.

Castura, J. C. (2009). Do panellists donkey vote in sensory choose-all-that-apply questions? In, *8th Pangborn Sensory Science Symposium, July 26-30*. Florence, Italy.

Castura, J. C., Antúnez, L., Giménez, A., & Ares, G. (2016). Temporal Check-All-That-Apply (TCATA): A novel dynamic method for characterizing products. *Food Quality and Preference, 47*, 79-90.

Clark, C. C., & Lawless, H. T. (1994). Limiting response alternatives in time-intensity scaling: an examination of the halo-dumping effect. *Chemical Senses, 19*(6), 583-594.

Cochran, W. G. (1950). The Comparison of Percentages in Matched Samples. *Biometrika, 37*(3/4), 256-266.

Coulon-Leroy, C., Symoneaux, R., Lawrence, G., Mehinagic, E., & Maitre, I. (2017). Mixed Profiling: A new tool of sensory analysis in a professional context. Application to wines. *Food Quality and Preference, 57*, 8-16.

Courcoux, P., Chaunier, L., Valle, G. D., Lourdin, D., & Séménou, M. (2005). Paired comparisons for the evaluation of crispness of cereal flakes by untrained assessors: correlation with descriptive analysis and acoustic measurements. *Journal of Chemometrics, 19*(3), 129-137.

Dairou, V., & Sieffermann, J. M. (2002). A Comparison of 14 Jams Characterized by Conventional Profile and a Quick Original Method, the Flash Profile. *Journal of Food Science, 67*(2), 826-834.

Danzart, M. (2009). *SSHA 3eme (Ed.), Evaluation sensorielle. Manuel méthodologique*. Paris: Lavoisier.

Depledt, F., & Sauvageot, F. (2002). Évaluation sensorielle des produits alimentaires. In, *Techniques de l'ingénieur Biochimie alimentaire, analyses et alimentation humaine*.

Dinnella, C., Masi, C., Zoboli, G., & Monteleone, E. (2012). Sensory functionality of extra-virgin olive oil in vegetable foods assessed by Temporal Dominance of Sensations and Descriptive Analysis. *Food Quality and Preference, 26*(2), 141-150.

Duizer, L. M., Bloom, K., & Findlay, C. J. (1996). Dual-attribute Time-intensity Measurement of Sweetness and Peppermint Perception of Chewing Gum. *Journal of Food Science, 61*(3), 636-638.

Dun-Rankin, P. (1983). *Scaling Methods*. Hillsdale, NJ: L. Erlbaum.

Escofier, B., & Pagès, J. (1994). Multiple factor analysis (AFMULT package). *Computational Statistics & Data Analysis, 18*(1), 121-140.

Foddy, W. (1993). *Constructing Questions for Interviews and Questionnaires: Theory and Practice in Social Research*. Cambridge: Cambridge University Press.

Gacula Jr, M., & Rutenbeck, S. (2006). Sample size in consumer test and descriptive analysis. *Journal of Sensory Studies, 21*(2), 129-145.

Galmarini, M. V., Symoneaux, R., Visalli, M., Zamora, M. C., & Schlich, P. (2016). Could Time–Intensity by a trained panel be replaced with a progressive profile done by consumers? A case on chewing-gum. *Food Quality and Preference, 48*, 274-282.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers, 26*(4), 381-386.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika, 40*(1), 33-51.

Green, B. G., Shaffer, G. S., & Gilmore, M. M. (1993). Derivation and evaluation of a semantic scale of oral sensation magnitude with apparent ratio properties. *Chemical Senses, 18*(6), 683-702.

Greenacre, M. (1988). Clustering the rows and columns of a contingency table. *Journal of Classification, 5*(1), 39-51.

Greenhoff, K., & MacFie, H. J. H. (1994). Preference mapping in practice. In, *H. J. H. MacFie & D. M. H. Thompson (Eds.), Measurement of food preferences*. London: Blackie Academic & Professionals.

Hanaei, F., Cuvelier, G., & Sieffermann, J. M. (2015). Consumer texture descriptions of a set of processed cheese. *Food Quality and Preference, 40*, 316-325.

Harshman, R. A., & Lundy, M. E. (1994). PARAFAC: Parallel factor analysis. *Computational Statistics & Data Analysis, 18*(1), 39-72.

Heymann, H., Machado, B., Torri, L., & Robinson, A. L. (2012). How many judges should one use for sensory descriptive analysis? *Journal of Sensory Studies, 27*(2), 111-122.

Holway, A. H., & Hurvich, L. M. (1937). Differential Gustatory Sensitivity to Salt. *The American Journal of Psychology, 49*(1), 37-48.

Hulin, W. S., & Katz, D. (1935). The Frois-Wittmann pictures of facial expression. *Journal of Experimental Psychology, 18*(4), 482-498.

Husson, F., Le Dien, S., & Pagès, J. (2001). Which value can be granted to sensory profiles given by consumers? Methodology and results. *Food Quality and Preference, 12*(5), 291-296.

Hutchings, S. C., Foster, K. D., Grigor, J. M. V., Bronlund, J. E., & Morgenstern, M. P. (2014). Temporal dominance of sensations: A comparison between younger and older subjects for the perception of food texture. *Food Quality and Preference, 31*, 106-115.

Israel, G. D., & Taylor, C. L. (1990). Can response order bias evaluations? *Evaluation and Program Planning, 13*(4), 365-371.

Jack, F. R., Piggott, J. R., & Paterson, A. (1994). Analysis of Textural Changes in Hard Cheese during Mastication by Progressive Profiling. *Journal of Food Science, 59*(3), 539-543.

Jaeger, S. R., Alcaire, F., Hunter, D. C., Jin, D., Castura, J. C., & Ares, G. (2018). Number of terms to use in temporal check-all-that-apply studies (TCATA and TCATA Fading) for sensory product characterization by consumers. *Food Quality and Preference, 64*, 154-159.

Jaeger, S. R., Beresford, M. K., Hunter, D. C., Alcaire, F., Castura, J. C., & Ares, G. (2017). Does a familiarization step influence results from a TCATA task? *Food Quality and Preference, 55*, 91-97.

Jaeger, S. R., Chheang, S. L., Jin, D., Roigard, C. M., & Ares, G. (2020). Check-all-that-apply (CATA) questions: Sensory term citation frequency reflects rated term intensity and applicability. *Food Quality and Preference, 86*, 103986.

Jellinek, G. (1964). Introduction to and Critical Review of Modern Methods of Sensory Analysis (odour, Taste and Flavour Evaluation) with Special Emphasis on Descriptive Sensory Analysis (flavour Profile Method). *Journal of Nutrition and Dietetics, 1*, 219-260.

Jones, L. V., Peryam, D. R., & Thurstone, L. L. (1955). Development of a scale for measuring soldiers' food preferences. *Journal of Food Science, 20*(5), 512-520.

Kelly, G. A. (1955). *The psychology of personal constructs*. New York: Norton.

Kim, I.-A., Hopkinson, A., van Hout, D., & Lee, H.-S. (2017). A novel two-step rating-based 'double-faced applicability' test. Part 1: Its performance in sample discrimination in comparison to simple one-step applicability rating. *Food Quality and Preference, 56*, 189-200.

Klayman, J., & Ha, Y.-w. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*(2), 211-228.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 107-118.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213-236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537-567.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response–order effects in survey measurement. *Public Opinion Quarterly, 51*(2), 201-219.

Kuesten, C., Bi, J., & Feng, Y. (2013). Exploring taffy product consumption experiences using a multi-attribute time–intensity (MATI) method. *Food Quality and Preference, 30*(2), 260-273.

Lagrange, V., & Norback, J. P. (1987). Product optimization and the Acceptor Set Size. *Journal of Sensory Studies, 2*(2), 119-136.

Lahne, J., Trubek, A. B., & Pelchat, M. L. (2014). Consumer sensory perception of cheese depends on context: A study using comment analysis and linear mixed models. *Food Quality and Preference, 32*, 184-197.

Lavit, C., Escoufier, Y., Sabatier, R., & Traissac, P. (1994). The ACT (STATIS method). *Computational Statistics & Data Analysis, 18*(1), 97-119.

Lawless, H. T. (1989). Exploration of Fragrance Categories and Ambiguous Odors Using Multidimensional-Scaling and Cluster-Analysis. *Chemical Senses, 14*(3), 349-360.

Lawless, H. T., & Heymann, H. (1999). *Sensory evaluation of food: Principles and practices*. New York: Kluwer Academic/Plenum Publishers.

Lawless, H. T., Sheng, N., & Knoops, S. S. C. P. (1995). Multidimensional-Scaling of Sorting Data Applied to Cheese Perception. *Food Quality and Preference, 6*(2), 91-98.

Lebart, L., & Salem, A. (1994). Statistique textuelle.

Lee, W. E., III, & Pangborn, R. M. (1986). Time-intensity: The temporal aspects of sensory perception. *Food Technology, 40*(11), 71-78.

Lenfant, F., Loret, C., Pineau, N., Hartmann, C., & Martin, N. (2009). Perception of oral food breakdown. The concept of sensory trajectory. *Appetite, 52*(3), 659-667.

Li, B., Hayes, J. E., & Ziegler, G. R. (2015). Maximizing overall liking results in a superior product to minimizing deviations from ideal ratings: An optimization case study with coffee-flavored milk. *Food Quality and Preference, 42*, 27-36.

Lu, J., Lin, C., Wang, W., Li, C., & Wang, H. (2013). String similarity measures and joins with synonyms. In, *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. New York, New York, USA: Association for Computing Machinery.

Luc, A., Lê, S., & Philippe, M. (2020). Nudging consumers for relevant data using Free JAR profiling: An application to product development. *Food Quality and Preference, 79*.

Malinvaud, E. (1964). *Méthodes statistiques de l'économétrie*. Paris: Dunod.

Mammasse, N., & Schlich, P. (2014). Adequate number of consumers in a liking test. Insights from resampling in seven studies. *Food Quality and Preference, 31*, 124-128.

McEwan, J. A. (1996). Preference Mapping for product optimization. In, *T. Naes & E. Risvik (Eds.), Multivariate analysis of data in sensory science*. New York: Elsevier.

Meilgaard, M., Civille, G. V., & Carr, B. T. (1991). *Sensory Evaluation Techniques*. Boca Raton, FL: CRC Press.

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.

Moskowitz, H. R. (1972). Subjective ideals and sensory optimization in evaluating perceptual dimensions in food. *Journal of Applied Psychology, 56*(1), 60-66.

Muñoz, A. M., & Civille, G. V. (1992). The Spectrum descriptive analysis method. In, *Am., ASTM Manual Series MNL 13, Manual On Descripive Analysis Testing*.

Navarro, G. (2001). A guided tour to approximate string matching. *33*(1 %J ACM Comput. Surv.), 31–88.

Nguyen, Q. C., Næs, T., & Varela, P. (2018). When the choice of the temporal method does make a difference: TCATA, TDS and TDS by modality for characterizing semi-solid foods. *Food Quality and Preference, 66*, 95-106.

Niedomysl, T., & Malmberg, B. (2009). Do open-ended survey questions on migration motives create coder variability problems? *Population, Space and Place, 15*(1), 79-87.

Pagès, J. (2005). Collection and analysis of perceived product inter-distances using multiple factor analysis: Application to the study of 10 white wines from the Loire Valley. *Food Quality and Preference, 16*(7), 642-649.

Payne, S. L. (1980). *The Art of Asking Questions*. Princeton N.J.: Princeton University Press.

Peltier, C., Mammasse, N., Visalli, M., Cordelle, S., & Schlich, P. (2018). Do we need to replicate in sensory profiling studies? *Food Quality and Preference, 63*, 129-134.

Perrin, L., Symoneaux, R., Maître, I., Asselin, C., Jourjon, F., & Pagès, J. (2008). Comparison of three sensory methods for use with the Napping® procedure: Case of ten wines from Loire valley. *Food Quality and Preference, 19*(1), 1-11.

Peryam, D. R., Pilgrim, F. J., & Peterson, M. S. (1954). *Food Acceptance Testing Methodology: A Symposium Sponsored by the Quartermaster Food and Container Institute for the Armed Forces, Quartermaster Research and Development Command, U.S. Army Quartermaster Corps [at The] Palmer House, Chicago, 8-9 October 1953*. Washington, D.C.: Advisory Board on Quartermaster Research and Development, Committee on Foods, Nat'l Academy of Sciences-National Research Council.

Pineau, N., Cordelle, S., Imbert, A., Rogeaux, M., & Schlich, P. (2003). Dominance temporelle des sensations - Codage et analyse d'un nouveau type de données sensorielles. In, *35èmes Journées de Statistiques, 2-6th June*. Lyon, France.

Pineau, N., de Bouillé, A. G., Lepage, M., Lenfant, F., Schlich, P., Martin, N., et al. (2012). Temporal Dominance of Sensations: What is a good attribute list? *Food Quality and Preference*, *26*(2), 159-165.

Pineau, N., Schlich, P., Cordelle, S., Mathonnière, C., Issanchou, S., Imbert, A., et al. (2009). Temporal Dominance of Sensations: Construction of the TDS curves and comparison with time–intensity. *Food Quality and Preference, 20*(6), 450-455.

Poirson, E., Petiot, J.-F., & Richard, F. (2010). A method for perceptual evaluation of products by naive subjects: Application to car engine sounds. *International Journal of Industrial Ergonomics, 40*(5), 504-516.

Popper, R. (2014). Use of Just-About-Right Scales in Consumer Research. In, *P. Varela & G. Ares (Eds), Novel techniques in sensory characterization and consumer profiling*. Boca Raton, FL (United States): CRC Press.

Qannari, E. M., Wakeling, I., Courcoux, P., & MacFie, H. J. H. (2000). Defining the underlying sensory dimensions. *Food Quality and Preference*, *11*(1), 151-154.

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

Ratinaud, P. (2014). IRaMuTeQ : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. In. France.

Reja, U., Manfreda, K., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. Close-ended Questions in Web Questionnaires. *Developments in Applied Statistics, 19*(1), 159.

Risvik, E., McEwan, J. A., Colwill, J. S., Rogers, R., & Lyon, D. H. (1994). Projective mapping: A tool for sensory analysis and consumer research. *Food Quality and Preference, 5*(4), 263-269.

Rodrigues, J. F., Souza, V. R. d., Lima, R. R., Carneiro, J. d. D. S., Nunes, C. A., & Pinheiro, A. C. M. (2016). Temporal dominance of sensations (TDS) panel behavior: A preliminary study with chocolate. *Food Quality and Preference, 54*, 51-57.

Rousseau, B. (2015). Sensory discrimination testing and consumer relevance. *Food Quality and Preference, 43*, 122-125.

Saporta, G. (2006). *Probabilités, analyse des données et statistique (2e édition révisée et augmentée)*. Paris: Editions Technip.

Schlich, P. (2017). Temporal Dominance of Sensations (TDS): a new deal for temporal sensory analysis. *Current Opinion in Food Science, 15*, 38-42.

Schlich, P., & McEwan, J. A. (1992). Cartographie des préférences. Un outil statistique pour l'industrie agro-alimentaire. *Science des Aliments, 12*, 339-355.

Schuman, H., & Presser, S. (1979). The Open and Closed Question. *American Sociological Review, 44*(5), 692-712.

Schuman, H., & Scott, J. (1987). Problems in the Use of Survey Questions to Measure Public Opinion. *Science, 236*(4804), 957.

Schwarz, N., & Hippler, H. J. (1991). Response alternatives: The impact of their choice and presentation order. In, *P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, & S. Sudman (Eds.) Measurement error in surveys*. Chichester: Wiley.

Sheatsley, P. B. (1983). Questionnaire construction and item writing. In, *P.H. Rossi, J.D. Wright, and A.B. Anderson (Eds.): Handbook of Survey Research*: Academic Press.

Silva, R. d. C. d. S. N. d., Minim, V. P. R., Silva, A. N. d., & Minim, L. A. (2014). Number of judges necessary for descriptive sensory tests. *Food Quality and Preference, 31*, 22-27.

Sjöström, L. B. (1954). The descriptive analysis of flavour. In, *D. R. Peryam, F. J. Pilgrim & M. S. Peterson, Food Acceptance Testing Methodology*. Chicago, IL: U.S. QUARTERMASTER FOOD AND CONTAINER INSTITUTE.

Stone, H., & Sidel, J. L. (1993). *Sensory evaluation practices*. California: Academic Press.

Stone, H., Sidel, J. L., Oliver, S., Woolsey, A., & Singleton, R. C. (1974). Sensory evaluation by quantitative descriptive analysis. *Food Technology*, *28*(11), 24-33.

Strigler, F., Touraille, C., Sauvageot, F., Barthélémy, J., Issanchou, S., & Pagès, J. (2009). Les épreuves discriminatives et descriptives. In, *Evaluation sensorielle : manuel méthodologique (3e édition)*. Paris: Lavoisier, Tec et Doc.

Sudman, S., & Bradburn, N. (1974). *Response Effects in Survey*.

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.

Teillet, E., Schlich, P., Urbano, C., Cordelle, S., & Guichard, E. (2010). Sensory methodologies and the taste of water. *Food Quality and Preference, 21*(8), 967-976.

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.

Thomas, A., Visalli, M., Cordelle, S., & Schlich, P. (2015). Temporal Drivers of Liking. *Food Quality and Preference, 40*, 365-375.

Thomson, D. M. H., & McEwan, J. A. (1988). An application of the repertory grid method to investigate consumer perceptions of foods. *Appetite, 10*(3), 181-193.

Thuillier, B. (2007). *Rôle du CO2 dans l'Appréciation Organoleptique des Champagnes – Expérimentation et Apports Méthodologiques*. Reims, France: Thèse de l'URCA.

Van Trijp, H. C. M., Punter, P. H., Mickartz, F., & Kruithof, L. (2007). The quest for the ideal product: Comparing different methods and approaches. *Food Quality and Preference, 18*(5), 729-740.

Varela, P., Antúnez, L., Carlehög, M., Alcaire, F., Castura, J. C., Berget, I., et al. (2018). What is dominance? An exploration of the concept in TDS tests with trained assessors and consumers. *Food Quality and Preference, 64*, 72-81.

Vidal, L., Ares, G., Hedderley, D. I., Meyners, M., & Jaeger, S. R. (2018). Comparison of rate-all-that-apply (RATA) and check-all-that-apply (CATA) questions across seven consumer studies. *Food Quality and Preference, 67*, 49-58.

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.

Visalli, M., Mahieu, B., Thomas, A., & Schlich, P. (2020a). Automated sentiment analysis of Free-Comment: An indirect liking measurement? *Food Quality and Preference, 82*.

Visalli, M., Mahieu, B., Thomas, A., & Schlich, P. (2020b). Concurrent vs. retrospective temporal data collection: Attack-evolution-finish as a simplification of Temporal Dominance of Sensations? *Food Quality and Preference, 85*.

Williams, A., Carr, B. T., & Popper, R. (2011). Exploring analysis options for check-all-that-apply (CATA) questions. In, *9th Rose-Marie Pangborn Sensory Science Symposium*. Toronto, ON, Canada.

Williams, A., & Langron, S. P. (1984). The use of free-choice profiling for the evaluation of commercial ports. *Journal of the Science of Food and Agriculture, 35*(5), 558-568.

Worch, T., Crine, A., Gruel, A., & Lê, S. (2014). Analysis and validation of the Ideal Profile Method: Application to a skin cream study. *Food Quality and Preference, 32*, 132-144.

Worch, T., Dooley, L., Meullenet, J.-F., & Punter, P. H. (2010). Comparison of PLS dummy variables and Fishbone method to determine optimal product characteristics from ideal profiles. *Food Quality and Preference, 21*(8), 1077-1087.

Worch, T., Lê, S., & Punter, P. (2010). How reliable are the consumers? Comparison of sensory profiles from consumers and experts. *Food Quality and Preference, 21*(3), 309-318.

Worch, T., Lê, S., Punter, P., & Pagès, J. (2012a). Assessment of the consistency of ideal profiles according to non-ideal data for IPM. *Food Quality and Preference, 24*(1), 99-110.

Worch, T., Lê, S., Punter, P., & Pagès, J. (2012b). Extension of the consistency of the data obtained with the Ideal Profile Method: Would the ideal products be more liked than the tested products? *Food Quality and Preference, 26*(1), 74-80.

Worch, T., Lê, S., Punter, P., & Pagès, J. (2013). Ideal Profile Method (IPM): The ins and outs. *Food Quality and Preference, 28*(1), 45-59.

Yzerbyt, V. Y., & Leyens, J.-P. (1991). Requesting information to form an impression: The influence of valence and confirmatory status. *Journal of Experimental Social Psychology, 27*(4), 337-356.

Züll, C. (2016). Open-Ended Questions. *GESIS Survey Guidelines*.

Appendix: R package “MultiResponseR”

Installing MultiResponseR:

```
> install.packages(“devtools”)
> devtools::install_github(“MahieuB/MultiResponseR”)
```

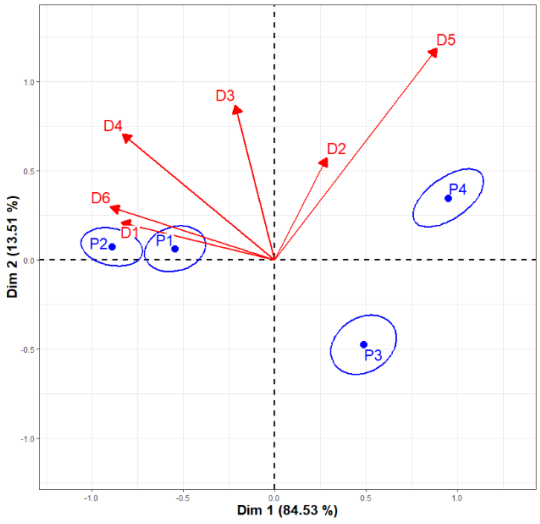
Using MultiResponseR:

```
> library(MultiResponseR)
> ?MultiResponseR
```

Then go to the “See Also” section.

Example of runs:

```
> data(milkchoc)
> dim.sig=sensory.mr.dimensionality.test(milkchoc)
> dim.sig
$dim.sig
[1] 3
$statistics
    Dim. 1     Dim. 2     Dim. 3
184.637642  28.556750   3.607564
$p.values
      Dim. 1       Dim. 2       Dim. 3
0.0004997501 0.0004997501 0.0024987506
> res=sensory.mrCA(milkchoc,nbaxes.sig=dim.sig$dim.sig)
> plt.mrCA(res)
```



```
> tab=sensory.mr.sig.cell(milkchoc,nbaxes.sig=dim.sig$dim.sig)
> plt.mr.sig.cell(tab)
```

	P1	P2	P3	P4
D1	68.57	74.29	27.14	18.57
D2	7.14	1.43	4.29	12.86
D3	37.14	34.29	12.86	31.43
D4	32.86	51.43	4.29	10.00
D5	7.14	2.86	15.71	50.00
D6	74.29	81.43	27.14	20.00

Documentation of the package:

Package 'MultiResponseR'

April 19, 2021

Title Analysis of multiple-response contingency data

Version 1.0.0

Description This package implements the multiple-response chi-square framework introduced in Mahieu, Schlich, Visalli, and Cardot (2021) for the analysis of contingency data obtained from multiple-response questionnaires. The multiple-response framework notably includes a multiple-response chi-square test and a multiple-response correspondence analysis. Two types of cases are distinguished: a general case where one subject contributes only to one category and the particular case of sensory data (notably Check-All-That-Apply and Free-Comment) where one subject contributes to several categories (products).

License GPL-3

Encoding UTF-8

LazyData false

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.1

Depends R (>= 4.0.0)

Imports stats,
FactoMineR,
graphics,
parallel,
doParallel,
foreach,
iterators,
candisc,
flextable,
officer,
abind,
ggplot2,
ggrepel

R topics documented:

`milkchoc` *Check-All-That-Apply data*

Description

Data coming from a Check-All-That-Apply experiment on milk chocolates. This dataset corresponds to the flavor data used as examples and described more precisely in Mahieu, Schlich, Visalli, and Cardot (2021)

Usage

```
data(milkchoc)
```

Format

An object of class `data.frame` with 280 rows and 8 columns.

References

Mahieu, B., Schlich, P., Visalli, M., & Cardot, H. (2021). A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data. Food Quality and Preference, 93.

Examples

```
data(milkchoc)
```

`mr.chisq.test` *Multiple-response chi-square test*

Description

Performs a multiple-response chi-square test as defined in Loughin and Scherer (1998) using random permutations to estimate the null distribution

Usage

```
mr.chisq.test(data, nperm = 2000, ncores = 2)
```

Arguments

data	A data.frame of observations in rows whose first column is a factor (the categories) and subsequent columns are binary numeric or integer, each column being a response option
nperm	Number of permuted datasets to estimate the distribution of the statistic under the null hypothesis. See details
ncores	Number of cores used to estimate the null distribution. Default is 2. See details

Details

- **nperm**: The distribution of the statistic under the null hypothesis of no associations between categories and response options is estimated using *nperm* datasets generated thanks to random permutations of the response vectors along observations. Note that this differs from the original proposition of Loughin and Scherer (1998) who used a parametric bootstrap to do so.
- **ncores**: The more cores are added in the process, the faster the results will be obtained. The number of available cores is accessible using `detectCores`. The parallel tasks are closed once the *nperm* datasets are generated.

Value

A list with the following elements:

statistic Observed multiple-response chi-square statistic

p.value p-value of the test

References

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

Mahieu, B., Schlich, P., Visalli, M., & Cardot, H. (2021). A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data. Food Quality and Preference, 93.

Examples

```
nb.obs=200
nb.response=5
nb.category=5
vec.category=paste("C",1:nb.category,sep="")
right=matrix(rbinom(nb.response*nb.obs,1,0.25),nb.obs,nb.response)
category=sample(vec.category,nb.obs,replace = TRUE)
dset=cbind.data.frame(category,right)
dset$category=as.factor(dset$category)

parallel::detectCores()

mr.chisq.test(dset)
```

`mr.dimensionality.test` *Multiple-response dimensionality test*

Description

Performs a multiple-response dimensionality test as defined in Mahieu, Schlich, Visalli, and Cardot (2021) using random permutations to estimate the null distribution

Usage

```
mr.dimensionality.test(data, nperm = 2000, alpha = 0.05, ncores = 2)
```

Arguments

`data`	A data.frame of observations in rows whose first column is a factor (the categories) and subsequent columns are binary numeric or integer, each column being a response option
`nperm`	Number of permuted datasets to estimate the distribution of the statistic under the null hypothesis. See details
`alpha`	The alpha risk of the test
`ncores`	Number of cores used to estimate the null distribution. Default is 2. See details

Details

- **nperm**: The distribution of the statistic under the null hypothesis of no associations between categories and response options is estimated using *nperm* datasets generated thanks to random permutations of the response vectors along observations.
- **ncores**: The more cores are added in the process, the faster the results will be obtained. The number of available cores is accessible using `detectCores`. The parallel tasks are closed once the *nperm* datasets are generated.

Value

A list with the following elements:

dim.sig The number of significant dimensions

statistics Observed multiple-response chi-square statistic of each dimension

p.values P-value of the test of each dimension adjusted for closed testing procedure

References

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

Mahieu, B., Schlich, P., Visalli, M., & Cardot, H. (2021). A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data. Food Quality and Preference, 93.

Examples

```
nb.obs=200
nb.response=5
nb.category=5
vec.category=paste("C",1:nb.category,sep="")
right=matrix(rbinom(nb.response*nb.obs,1,0.25),nb.obs,nb.response)
category=sample(vec.category,nb.obs,replace = TRUE)
dset=cbind.data.frame(category,right)
dset$category=as.factor(dset$category)

parallel::detectCores()

mr.dimensionality.test(dset)
```

`mr.sig.cell` *Multiple-response tests per cell*

Description

This function performs for each pair of category and response option a multiple-response hypergeometric test as defined in Mahieu, Schlich, Visalli, and Cardot (2021) using random hypergeometric samplings to estimate the null distribution

Usage

```
mr.sig.cell(
  data,
  nsample = 2000,
  nbaxes.sig = Inf,
  two.sided = FALSE,
  ncores = 2
)
```

Arguments

`data`	A data.frame of observations in rows whose first column is a factor (the categories) and subsequent columns are binary numeric or integer, each column being a response option
`nsample`	Number of randomly sampled datasets to estimate the distribution of the value under the null hypothesis. See details
`nbaxes.sig`	The number of significant axes retuned by `mr.dimensionality.test`. By default, all axes are considered significant. See details
`two.sided`	Logical. Should the tests be two-sided or not? By default, the tests are performed with a one-sided greater alternative hypothesis
`ncores`	Number of cores used to estimate the null distribution. Default is 2. See details

Details

- **nsample**: The distribution of the value under the null hypothesis of no associations between categories and response options is estimated using *nsample* datasets generated thanks to random hypergeometric samplings of the response vectors along observations.
- **nbaxes.sig**: If *nbaxes.sig* is lower than the total number of axes then the tests are performed on the derived contingency table corresponding to significant axes (Mahieu, Schlich, Visalli, & Cardot, 2021). This table is obtained by using the reconstitution formula of MR-CA on the first *nbaxes.sig* axes.
- **ncores**: The more cores are added in the process, the faster the results will be obtained. The number of available cores is accessible using `detectCores`. The parallel tasks are closed once the *nsample* datasets are generated.

Value

A list with the following elements:

original.cont Observed number of times each category chosen each response option

percent.cont Within each category, percentage of observations where the response options were chosen

null.cont Expected number of times each category chosen each response option under the null hypothesis

p.values P-values of the tests per cell

derived.cont The derived contingency table corresponding to *nbaxes.sig* axes

percent.derived.cont Within each category, percentage of observations where the response options were chosen in the derived contingency table corresponding to *nbaxes.sig* axes

References

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

Mahieu, B., Schlich, P., Visalli, M., & Cardot, H. (2021). A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data. Food Quality and Preference, 93.

Examples

```
nb.obs=200
nb.response=5
nb.category=5
vec.category=paste("C",1:nb.category,sep="")
right=matrix(rbinom(nb.response*nb.obs,1,0.25),nb.obs,nb.response)
category=sample(vec.category,nb.obs,replace = TRUE)
dset=cbind.data.frame(category,right)
dset$category=as.factor(dset$category)

parallel::detectCores()

res=mr.sig.cell(dset)

plt.mr.sig.cell(res)
```

mrCA *Multiple-response Correspondence Analysis (MR-CA)*

Description

This functions performs a multiple-response Correspondence Analysis (MR-CA) as defined in Mahieu, Schlich, Visalli, and Cardot (2021)

Usage

```
mrCA(
  data,
  proj.row = NULL,
  proj.row.obs = NULL,
  proj.col = NULL,
  ellipse = FALSE,
  nboot = 2000,
  nbaxes.sig = Inf,
  ncores = 2
)
```

Arguments

data	A data.frame of observations in rows whose first column is a factor (the categories) and subsequent columns are binary numeric or integer, each column being a response option
proj.row	Optional. A contingency table with new categories to be projected as supplementary rows within the MR-CA space in rows and the same response options as data in columns
proj.row.obs	A numeric vector whose length equals nrow(proj.row) and giving the number of observations within each projected rows. Useless if proj.row=NULL
proj.col	Optional. A contingency table with new response options to be projected as supplementary columns within the MR-CA space in columns and the same categories as data in rows
ellipse	Logical. Does confidence ellipses for the categories should be computed? Default is FALSE. See details
nboot	Number of virtual datasets used in the total bootsrap procedure. Useless when ellipse=FALSE. See details
nbaxes.sig	The number of significant axes retuned by mr.dimensionality.test. By default, all axes are considered significant. Useless when ellipse=FALSE. See details
ncores	Number of cores used to generate the virtual datasets Default is 2. Useless when ellipse=FALSE. See details

Details

- **ellipse**: When ellipse=TRUE, confidence ellipses for the categories are computed using a total bootstrap procedure (Cadoret & Husson, 2013). **nboot** virtual datasets are generated by randomly sample with replacement response option within each category. A MR-CA is then

performed on these virtual dataset and the virtual configurations are adjusted on the actual configuration using Procrustes rotations accounting for **nbaxes.sig** axes (Mahieu, Schlich, Visalli, & Cardot, 2021). Finally, for each category, a confidence ellipse is constructed using the position of its bootstrap replicates. The ellipses are plotted when using `plt.mrCA` Pairwise total bootstrap tests (Mahieu, Visalli, Thomas, & Schlich, 2020) are also performed between the categories

- **ncores**: The more cores are added in the process, the faster the results will be obtained. The number of available cores is accessible using `detectCores`. The parallel tasks are closed once the *nboot* datasets are generated.

Value

A list with the following elements:

eigen Eigenvalues and their corresponding percentages of inertia

row.coord Rows coordinates

col.coord Columns coordinates

proj.row.coord Projected rows coordinates

proj.col.coord Projected columns coordinates

svd Results of the singular value decomposition

bootstrap.replicate.coord Coordinates of the rotated bootstrap replicates

total.bootstrap.test.pvalues P-values of the pairwise total bootstrap tests

References

Mahieu, B., Schlich, P., Visalli, M., & Cardot, H. (2021). A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data. Food Quality and Preference, 93.

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

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

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

Examples

```
nb.obs=200
nb.response=5
nb.category=5
vec.category=paste("C",1:nb.category,sep="")
right=matrix(rbinom(nb.response*nb.obs,1,0.25),nb.obs,nb.response)
category=sample(vec.category,nb.obs,replace = TRUE)
dset=cbind.data.frame(category,right)
dset$category=as.factor(dset$category)



res=mrCA(dset)

plt.mrCA(res)
```

`plt.mr.sig.cell` *Plot significant cells*

Description

This function plots the results coming from `sensory.mr.sig.cell` or `mr.sig.cell`

Usage

```
plt.mr.sig.cell(
  res,
  alpha = 0.05,
  choice = "percent.derived.cont",
  col.greater = "green3",
  col.lower = "orangered"
)
```

Arguments

`res`	A list returned by `sensory.mr.sig.cell` or `mr.sig.cell`
`alpha`	The alpha risk to consider the tests as significant
`choice`	Which table from *res* should be plotted? Default is percent.derived.cont
`col.greater`	The color used to highlight significant positive associations
`col.lower`	The color used to highlight significant negative associations

Value

A table with significant cells highlighted

Examples

```
# non-sensory example
nb.obs=200
nb.response=5
nb.category=5
vec.category=paste("C",1:nb.category,sep="")
right=matrix(rbinom(nb.response*nb.obs,1,0.25),nb.obs,nb.response)
category=sample(vec.category,nb.obs,replace = TRUE)
dset=cbind.data.frame(category,right)
dset$category=as.factor(dset$category)

parallel::detectCores()

res=mr.sig.cell(dset)

plt.mr.sig.cell(res)

# sensory example
data(milkchoc)

parallel::detectCores()
```

```
dim.sig=sensory.mr.dimensionality.test(milkchoc)$dim.sig

res=sensory.mr.sig.cell(milkchoc,nbaxes.sig=dim.sig)

plt.mr.sig.cell(res)
```

`plt.mrCA` *Plot factor plan resulting from multiple-response Correspondence Analysis (MR-CA)*

Description

This function plots the results coming from `sensory.mrCA` or `mrCA`

Usage

```
plt.mrCA(
  res,
  axes = c(1, 2),
  alpha.total.bootstrap.test = 0.05,
  alpha.ellipse = alpha.total.bootstrap.test,
  select.desc.rep = rownames(res$col.coord),
  rev.x = FALSE,
  rev.y = FALSE,
  size.points = 3.5,
  size.lab = 6,
  expansion = 1.25,
  title = NULL
)
```

Arguments

`res`	A list returned by `sensory.mrCA` or `mrCA`
`axes`	Which dimensions of the MR-CA should be plotted?
`alpha.total.bootstrap.test`	The alpha risk of the total bootstrap tests. Only useful if the MR-CA was computed using `sensory.mrCA` or `mrCA` and ellipse=TRUE. See details
`alpha.ellipse`	The alpha risk of the confidence ellipses. Only useful if the MR-CA was computed using `sensory.mrCA` or `mrCA` and ellipse=TRUE
`select.desc.rep`	A character vector specifying the descriptors/response options to plot. By default, all descriptors/response options are plotted
`rev.x`	Should the horizontal plotted dimension be reversed? Useful in case of map comparisons to align products/categories
`rev.y`	Should the vertical plotted dimension be reversed? Useful in case of map comparisons to align products/categories
`size.points`	The size of the points used to represent the products/categories on the map
`size.lab`	The size of the label on the map
`expansion`	The factor of expansion applied to descriptors/response options coordinates to increase readability
`title`	An optional title to be added to the plot

Details

- **alpha.total.bootstrap.test**: Products/categories non-significantly different at the alpha risk of *alpha.total.bootstrap.test* according to the total bootstrap test are linked by a line on the plot. If these links are not required, *alpha.total.bootstrap.test* can be set to 1

Value

A MR-CA factor map

Examples

```
# non-sensory example
nb.obs=200
nb.response=5
nb.category=5
vec.category=paste("C",1:nb.category,sep="")
right=matrix(rbinom(nb.response*nb.obs,1,0.25),nb.obs,nb.response)
category=sample(vec.category,nb.obs,replace = TRUE)
dset=cbind.data.frame(category,right)
dset$category=as.factor(dset$category)

parallel::detectCores()

res=mrCA(dset)

plt.mrCA(res)

# sensory example
data(milkchoc)

parallel::detectCores()

dim.sig=sensory.mr.dimensionality.test(milkchoc)$dim.sig

res=sensory.mrCA(milkchoc,nbaxes.sig=dim.sig)

plt.mrCA(res)
```

---

sensory.mr.dimensionality.test

Multiple-response dimensionality test for sensory data

Description

Performs a multiple-response dimensionality test as defined in Mahieu, Schlich, Visalli, and Cardot (2021) using random permutations to estimate the null distribution. The difference with `mr.dimensionality.test` is that random permutations are performed within subjects rather than along all evaluations

Usage

```
sensory.mr.dimensionality.test(data, nperm = 2000, alpha = 0.05, ncores = 2)
```

Arguments

`data`	A data.frame of evaluations in rows whose first two columns are factors (subject and product) and subsequent columns are binary numeric or integer, each column being a descriptor
`nperm`	Number of permuted datasets to estimate the distribution of the statistic under the null hypothesis. See details
`alpha`	The alpha risk of the test
`ncores`	Number of cores used to estimate the null distribution. Default is 2. See details

Details

- **nperm**: The distribution of the statistic under the null hypothesis of no associations between products and descriptors is estimated using *nperm* datasets generated thanks to random permutations of the response vectors along products within subjects.
- **ncores**: The more cores are added in the process, the faster the results will be obtained. The number of available cores is accessible using `detectCores`. The parallel tasks are closed once the *nperm* datasets are generated.

Value

A list with the following elements:

dim.sig The number of significant dimensions

statistics Observed multiple-response chi-square statistic of each dimension

p.values P-value of the test of each dimension adjusted for closed testing procedure

References

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

Mahieu, B., Schlich, P., Visalli, M., & Cardot, H. (2021). A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data. Food Quality and Preference, 93.

Examples

```
data(milkchoc)

parallel::detectCores()

sensory.mr.dimensionality.test(milkchoc)
```

`sensory.mr.sig.cell` *Multiple-response tests per cell for sensory data*

Description

This function performs for each pair of product and descriptor a multiple-response hypergeometric test as defined in Mahieu, Schlich, Visalli, and Cardot (2021) using random hypergeometric samplings to estimate the null distribution. The difference with `mr.sig.cell` is that random hypergeometric samplings are performed taking into account the subject structure of sensory data in `sensory.mr.sig.cell`

Usage

```
sensory.mr.sig.cell(
  data,
  nsample = 2000,
  nbaxes.sig = Inf,
  two.sided = FALSE,
  ncores = 2
)
```

Arguments

`data`	A data.frame of evaluations in rows whose first two columns are factors (subject and product) and subsequent columns are binary numeric or integer, each column being a descriptor
`nsample`	Number of randomly sampled datasets to estimate the distribution of the value under the null hypothesis. See details
`nbaxes.sig`	The number of significant axes retuned by `sensory.mr.dimensionality.test`. By default, all axes are considered significant. See details
`two.sided`	Logical. Should the tests be two-sided or not? By default, the tests are performed with a one-sided greater alternative hypothesis
`ncores`	Number of cores used to estimate the null distribution. Default is 2. See details

Details

- **nsample**: The distribution of the value under the null hypothesis of no associations between products and descriptors is estimated using *nsample* datasets generated thanks to random hypergeometric samplings of the response vectors along products within subjects.
- **nbaxes.sig**: If *nbaxes.sig* is lower than the total number of axes then the tests are performed on the derived contingency table corresponding to significant axes (Mahieu, Schlich, Visalli, & Cardot, 2021) This table is obtained by using the reconstitution formula of MR-CA on the first *nbaxes.sig* axes.
- **ncores**: The more cores are added in the process, the faster the results will be obtained. The number of available cores is accessible using `detectCores`. The parallel tasks are closed once the *nsample* datasets are generated.

Value

A list with the following elements:

original.cont Observed number of times each product was described by each descriptor

percent.cont For each product, percentage of evaluations where each descriptor was cited for this product

null.cont Expected number of times each product was described by each descriptor under the null hypothesis

p.values P-values of the tests per cell

derived.cont The derived contingency table corresponding to *nbaxes.sig* axes

percent.derived.cont For each product, percentage of evaluations where each descriptor was cited for this product in the derived contingency table corresponding to *nbaxes.sig* axes

References

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

Mahieu, B., Schlich, P., Visalli, M., & Cardot, H. (2021). A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data. Food Quality and Preference, 93.

Examples

```
data(milkchoc)

parallel::detectCores()

dim.sig=sensory.mr.dimensionality.test(milkchoc)$dim.sig

res=sensory.mr.sig.cell(milkchoc,nbaxes.sig=dim.sig)

plt.mr.sig.cell(res)
```

`sensory.mrCA` *Multiple-response Correspondence Analysis (MR-CA) for sensory data*

Description

This function performs the MR-CA of the data as well as the total bootstrap procedure (Cadoret & Husson, 2013) and the pairwise total bootstrap tests (Mahieu, Visalli, Thomas, & Schlich, 2020). The difference with `mrCA` used with ellipse=TRUE is that the total bootstrap procedure takes into account the subject structure of sensory data in `sensory.mrCA`

Usage

```
sensory.mrCA(data, nboot = 2000, nbaxes.sig = Inf, ncores = 2)
```

Arguments

data	A data.frame of evaluations in rows whose first two columns are factors (subject and product) and subsequent columns are binary numeric or integer, each column being a descriptor
nboot	The number of bootstrapped panel of the total bootstrap procedure
nbaxes.sig	The number of significant axes retuned by sensory.mr.dimensionality.test. By default, all axes are considered significant. See details
ncores	Number of cores used to generate the virtual panels. Default is 2. See details

Details

- **nbaxes.sig**: The number of significant axes determines the number of axes accounted for while performing the Procrustes rotations of the total bootstrap procedure (Mahieu, Schlich, Visalli, & Cardot, 2021). These same axes are accounted for the pairwise total bootstrap tests.
- **ncores**: The more cores are added in the process, the faster the results will be obtained. The number of available cores is accessible using detectCores. The parallel tasks are closed once the *nboot* datasets are generated.

Value

A list with the following elements:

eigen Eigenvalues of the MR-CA and their corresponding percentages of inertia

row.coord Products coordinates

col.coord Descriptors coordinates

svd Results of the singular value decomposition

bootstrap.replicate.coord Coordinates of the rotated bootstrap replicates

total.bootstrap.test.pvalues P-values of the pairwise total bootstrap tests

References

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

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

Mahieu, B., Schlich, P., Visalli, M., & Cardot, H. (2021). A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data. Food Quality and Preference, 93.

Examples

```
data(milkchoc)

parallel::detectCores()

dim.sig=sensory.mr.dimensionality.test(milkchoc)$dim.sig

res=sensory.mrCA(milkchoc,nbaxes.sig=dim.sig)

plt.mrCA(res)
```

`sensory.overall.analysis` *Overall analysis of multiple-response sensory data using the multiple-response chi-square framework introduced in Mahieu, Schlich, Visalli, and Cardot (2021)*

Description

Successively performs `sensory.mr.dimensionality.test`, `sensory.mrCA` and `sensory.mr.sig.cell`

Usage

```
sensory.overall.analysis(
  data,
  nMC = 2000,
  alpha = 0.05,
  cell.two.sided = FALSE,
  ncores = 2
)
```

Arguments

`data`	A data.frame of evaluations in rows whose first two columns are factors (subject and product) and subsequent columns are binary numeric or integer, each column being a descriptor
`nMC`	Number of Monte-Carlo simulations to consider at each step of the overall analysis
`alpha`	The alpha risk to consider at each step of the overall analysis
`cell.two.sided`	Logical. Should the multiple-response tests per cell be two-sided or not? By default, the tests are performed with a one-sided greater alternative hypothesis
`ncores`	Number of cores used in the Monte-Carlo simulations. Default is 2. See details

Details

- **ncores**: The more cores are added in the process, the faster the results will be obtained. The number of available cores is accessible using `detectCores`. The parallel tasks are closed once the simulations are over.

Value

The first MR-CA factor map and the percent.derived.cont table with significant cells highlighted

References

Mahieu, B., Schlich, P., Visalli, M., & Cardot, H. (2021). A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data. Food Quality and Preference, 93.

Examples

```
data(milkchoc)

parallel::detectCores()

sensory.overall.analysis(milkchoc)
```

Index

Abstract

Free-Comment (FC) consists in panelists describing the products using their own terms. Despite its benefits, notably the circumvention of limitations inherent to pre-established lists of sensory descriptors, FC remains rarely used because its performances are not well documented and its analyses and range of application remain limited. This thesis aims to overpass these limitations, highlighting the benefits and the potency of FC and thus put it in the spotlight for sensory analysis with consumers.

For the pretreatment of FC data, a semi-automatized procedure is proposed. It enables the practitioners to extract an *a posteriori* list of sensory descriptors with a compromise between minimizing the loss of information and maximizing the quickness of the pretreatment. For the statistical analysis of FC data, operating in the significant subspace of product by sensory descriptor dependences is proposed together with the multiple-response chi-square framework that better takes into account the structure of the pretreated data than the usual chi-square framework. These analyses have been implemented into a R-package downloadable from GitHub.

The performances of FC have been compared to those of Check-All-That-Apply (CATA), the most popular method for descriptive sensory analysis with consumers. Two performance criteria have been investigated: the discrimination power and the stability of the product characterization. Regarding both criteria, FC turned out to perform as well as CATA, if not better.

Two extensions of FC are proposed. The first one, Free-Comment Attack-Evolution-Finish (FC-AEF), directs the descriptions towards the temporal aspect of the sensory perception. The second one, Ideal-Free-Comment (IFC) paired with liking scoring, identifies the drivers of liking and characterizes the ideal product thanks to FC. An application of these two methods was carried out, demonstrating their ability to fulfill their aims.

Overall, this work demonstrated the potency and the versatility of the FC method. It opens new perspectives for sensory analysis with consumers and it should promote a larger use of FC in that field.

Keywords: Open-ended questions; Free-Comment; Sensometrics; Sensory analysis; Consumer studies