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, Michel Visalli, Arnaud Thomas, Pascal Schlich

Title

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

Authors

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.

Corresponding author:

Benjamin Mahieu; Pascal Schlich

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

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

Highlights

- Stability of product configurations was investigated
- Stability of product by descriptor associations was investigated
- Free-Comment outputs were slightly more stable than Check-All-That-Apply ones
- Stability was strongly dependent on the size of product differences
- Product configurations were more stable than product by descriptor associations

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.

Keywords

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

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.

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). 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).

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

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

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

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

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

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 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 that 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).

The previous statements worth being nuanced by two points. 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 that was previously reported for CATA, RATA, 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 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

444 descriptive analysis (Gacula Jr & Rutenbeck, 2006; Heymann, Machado, Torri, &
445 Robinson, 2012; Silva, Minim, Silva, & Minim, 2014). This effect of the size of product
446 differences affected the stability of both FC and CATA in the same direction and with
447 the same magnitude.

448 For both FC and CATA, the product configurations were more stable than the joint
449 product by descriptor configurations, themselves being more stable than the product
450 by descriptor significant associations. This suggests that the more an aspect of the
451 outputs is demanding, the less it is stable. The product configurations are relatively
452 stable because they are driven by intrinsic differences between the products and do
453 not depend on how these intrinsic differences are transcribed and/or verbalized. This
454 is supported by several studies that compared two or more consumer sensory methods
455 and observed that they provided similar product configurations (Ares, Bruzzone, et al.,
456 2014; Fleming, Ziegler, & Hayes, 2015; Oppermann et al., 2017; Reinbach, Giacalone,
457 Ribeiro, Bredie, & Frøst, 2014). The joint product by descriptor configurations is less
458 stable than the product configuration because identifying differences is easier than
459 explicitly verbalizing them. However, the joint product by descriptor configurations is
460 still relatively stable because the big picture of each joint product by descriptor
461 configuration is likely to be recovered across repeated experiments. The product by
462 descriptor significant associations is at best moderately stable because they require
463 the intrinsic product differences to be verbalized significantly with the same descriptors
464 across repeated experiments, which is the most demanding aspect of the outputs.

465 4.2. Recommendations

466 When the investigated product space is characterized by large differences between
467 the products, 60 consumers enable to guarantee at least a medium stability of FC and
468 CATA outputs, which is in line with previous results concerning CATA (Ares, Tárrega,
469 et al., 2014). When differences between the products are more subtle, 60 consumers
470 enable to guarantee at least a medium stability of the product configurations and the
471 joint product by descriptor configurations for both FC and CATA but do not guarantee
472 stable product by descriptor significant associations. Future studies need to be
473 conducted to investigate the number of consumers necessary to obtain good stability
474 of the product by descriptor significant associations when working with products having
475 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., 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., 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., et al. (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*, 751-760.

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 Jr, 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*(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., 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*.

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*(3), 257-265.

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.

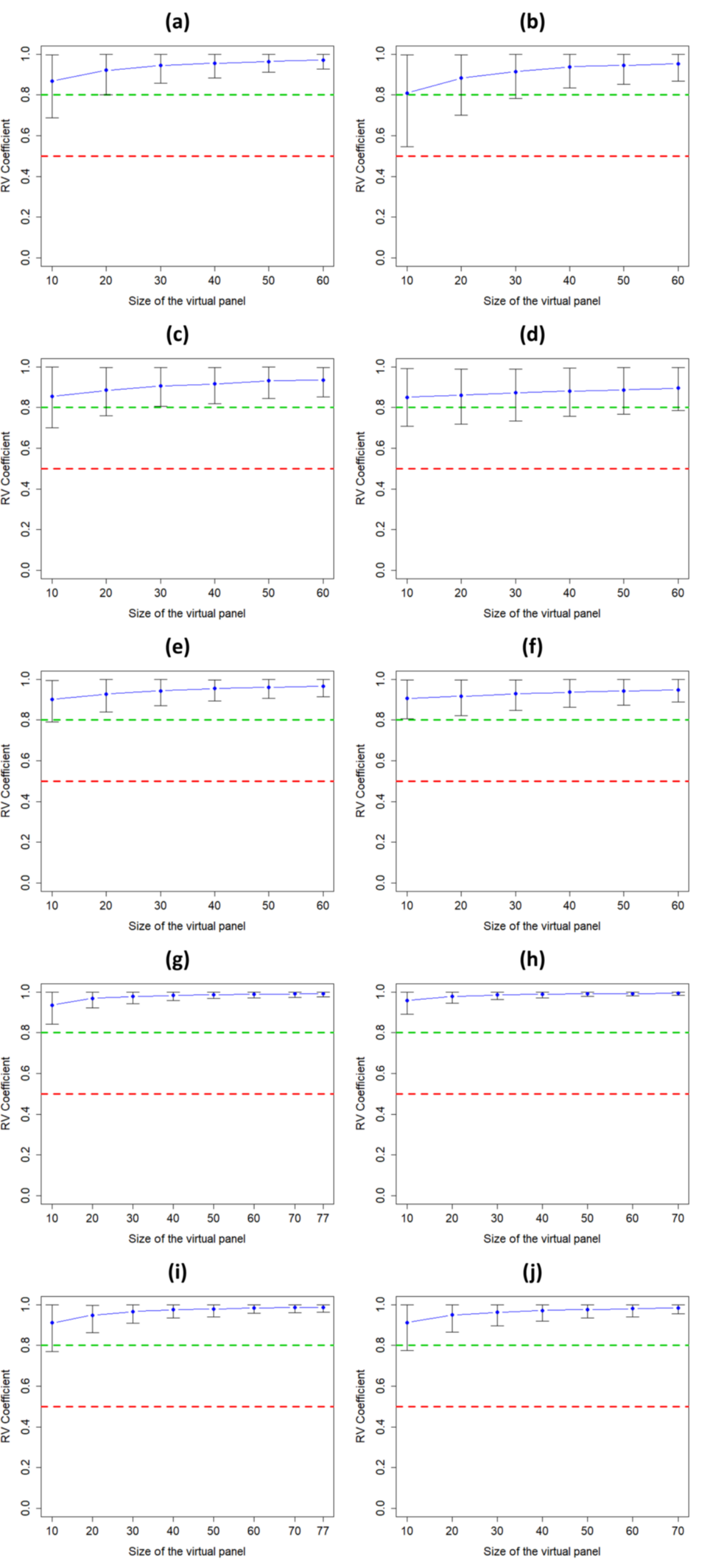
Valentin, D., Chollet, S., Lelièvre, M., & Abdi, H. (2012). Quick and dirty but still pretty good: a review of new descriptive methods in food science. *International Journal of Food Science & Technology, 47*(8), 1563-1578.

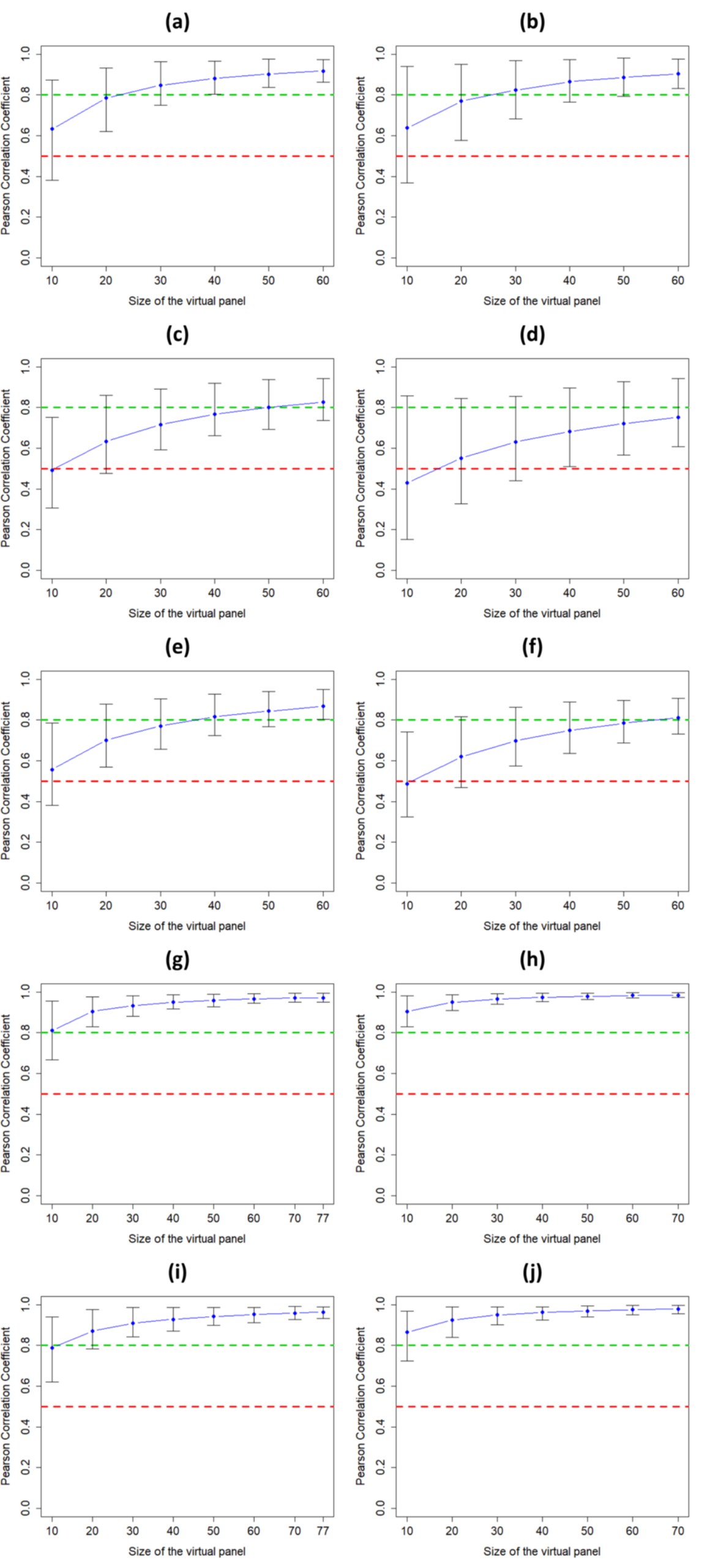
602 Varela, P., & Ares, G. (2012). Sensory profiling, the blurred line between sensory and
603 consumer science. A review of novel methods for product characterization.
604 *Food Research International, 48*(2), 893-908.
605 Vidal, L., Cadena, R. S., Antúnez, L., Giménez, A., Varela, P., & Ares, G. (2014).
606 Stability of sample configurations from projective mapping: How many
607 consumers are necessary? *Food Quality and Preference, 34*, 79-87.
608 Vidal, L., Tárrega, A., Antúnez, L., Ares, G., & Jaeger, S. R. (2015). Comparison of
609 Correspondence Analysis based on Hellinger and chi-square distances to
610 obtain sensory spaces from check-all-that-apply (CATA) questions. *Food*
611 *Quality and Preference, 43*, 106-112.

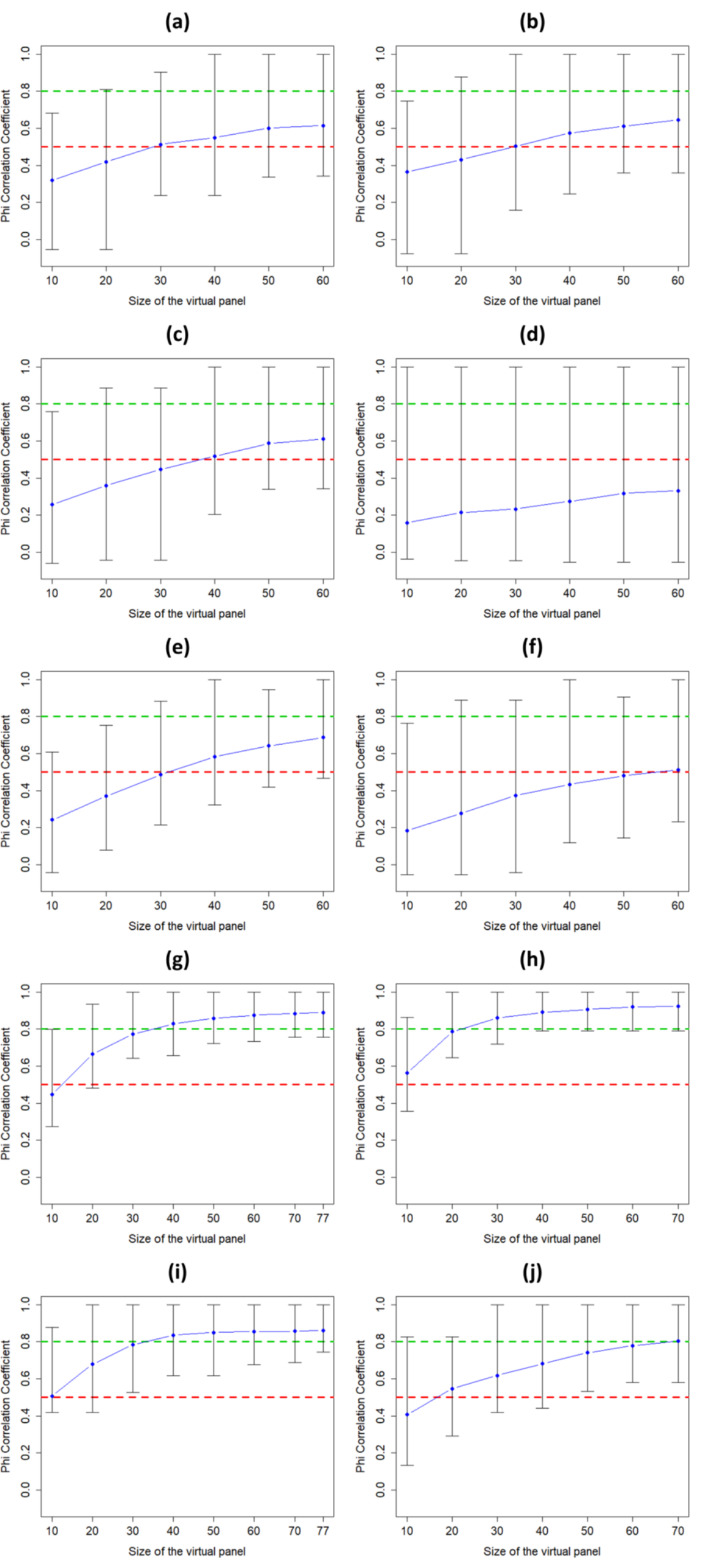
Fig. 1: Mean of the distribution of the RV coefficients between the actual and the virtual product configurations as a function of the virtual panel size for (a) FC-Wine-Vis, (b) CATA-Wine-Vis, (c) FC-Wine-Olf, (d) CATA-Wine-Olf, (e) FC-Wine-Gus, (f) CATA-Wine-Gus, (g) FC-Choc-Tex, (h) CATA-Choc-Tex, (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.

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.

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.


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
RV Coefficient
Size of the virtual panel
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
60
70
77


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
Pearson Correlation Coefficient
Size of the virtual panel
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
60
70
77


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
Phi Correlation Coefficient
Size of the virtual panel
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
60
70
77