



HAL
open science

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

► **To cite this version:**

Benjamin Mahieu, Michel Visalli, Arnaud Thomas, Pascal Schlich. 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, 2021, 90, pp.104159. 10.1016/j.foodqual.2020.104159 . hal-03191403

HAL Id: hal-03191403

<https://hal.inrae.fr/hal-03191403>

Submitted on 2 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

1 Title

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

4 Authors

5 Benjamin Mahieu^a, Michel Visalli^a, Arnaud Thomas^b, Pascal Schlich^a

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

8 ^bSensoStat, Dijon, France.

9 Corresponding author:

10 Benjamin Mahieu; Pascal Schlich

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

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

14 Highlights

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

22 Abstract

23 Free-Comment (FC), as a response to open-ended questions, enables a word-based
24 sensory description and discrimination of sets of products. The stability of FC outputs
25 has never been investigated and is the purpose of the present paper. Since Check-All-
26 That-Apply (CATA) is the most popular method for the word-based sensory description

27 of products with consumers, the stability of FC was compared to that of CATA
28 [performed on the same products](#). Four red wines and four milk chocolates were
29 evaluated according to different sensory modalities by groups of consumers following
30 either an FC or a CATA protocol. The stability of the product configurations and the
31 product by descriptor associations were investigated. FC outputs were slightly more
32 stable than CATA ones. Sixty consumers enable to guarantee medium stability, if not
33 good, of FC and CATA outputs when the investigated product space is characterized
34 by large differences between the products. The minimum number of consumers to
35 obtain stable results was strongly dependent on the size of the differences between
36 the products, which suggests that if *a priori* knowledge on the size of the differences
37 between the investigated products is available, it must drive the decision of the number
38 of consumers to include in the study rather than relying on an absolute rule. For both
39 FC and CATA, the product configurations were more easily stable in terms of numbers
40 of consumers than the product by descriptor associations. Investigating the stability of
41 the product by descriptor associations *a posteriori* is recommended for future FC and
42 CATA studies.

43 **Keywords**

- 44 - Open-ended questions
- 45 - Stability
- 46 - Sensory method comparison
- 47 - Consumer study

48 **1. Introduction**

49 Free-Comment (FC) ([ten Kleij & Musters, 2003](#)), as a response to open-ended
50 questions, is a sensory method that enables collecting word-based sensory
51 descriptions of a set of products without a predefined list of descriptors. For each
52 evaluated product, consumers are asked to describe the product in their own words
53 ([Ares, Giménez, Barreiro, & Gámbaro, 2010](#); [Hanaei, Cuvelier, & Sieffermann, 2015](#);
54 [Lahne, Trubek, & Pelchat, 2014](#); [Luc, Lê, & Philippe, 2020](#); [Mahieu, Visalli, Thomas,](#)
55 [& Schlich, 2020](#); [Symoneaux, Galmarini, & Mehinagic, 2012](#); [ten Kleij & Musters,](#)
56 [2003](#)). FC has already proven itself an efficient method in characterizing and
57 discriminating sets of products both with consumers and experts ([Lahne et al., 2014](#);

58 [Lawrence et al., 2013; ten Kleij & Musters, 2003](#)) even out of the lab ([Mahieu et al.,](#)
59 [2020](#)).

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

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

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

82 For consumer-oriented sensory methods, gathering a large number of different
83 experiments conducted under similar experimental settings with different panel sizes
84 is nearly impossible for practical limitations ([Ares, Tárrega, et al., 2014](#)). Thus, the
85 stability of the outputs is often evaluated internally, rather than externally, using
86 bootstrap resampling of an actual panel that performed a study in the experimental
87 settings under interest ([Ares, Bruzzone, et al., 2014; Ares, Tárrega, et al., 2014;](#)
88 [Blancher, Clavier, Egoroff, Duineveld, & Parcon, 2012; Cadena et al., 2014;](#)
89 [Mammasse & Schlich, 2014; Vidal et al., 2014; Vidal, Tárrega, Antúnez, Ares, &](#)
90 [Jaeger, 2015](#)). This procedure enables to generate a large number of virtual panels of

91 different sizes that simulate repeated experiments under similar experimental settings.
92 The outputs obtained from the actual panel are considered [as a benchmark to which](#)
93 [those of the virtual panels are compared](#).

94 Depending on the sensory method under investigation, different aspects of the outputs
95 are compared between the actual and the virtual panels. The product configurations
96 between the actual and the virtual panels were compared in every aforementioned
97 study using the RV coefficient ([Escoufier, 1973; Robert & Escoufier, 1976](#)). For word-
98 based sensory methods, the descriptor configurations were also compared using the
99 RV coefficient ([Ares, Bruzzone, et al., 2014; Ares, Tárrega, et al., 2014; Vidal et al.,](#)
100 [2015](#)). However, the descriptor configurations are usually not interpreted for
101 themselves but rather together with the product configurations to characterize the
102 product space. Thus, investigating the stability of the product by descriptor
103 associations rather than the stability of the descriptor configurations seems to be more
104 in line with common practices.

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

114 **2. Material and methods**

115 **2.1. Datasets**

116 The information concerning the datasets used in this paper and provided across the
117 material and methods section are summarized in [Table 2](#).

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

119 **2.1.1. First study: red wines**

120 The datasets of this study are the same from [Mahieu et al. \(2020\)](#).

121 *2.1.1.1. Participants*

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

131 *2.1.1.2. Products*

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

134 *2.1.1.3. FC task and datasets*

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

- 138 - Visual: “Describe the visual characteristics of the wine”
- 139 - Olfactory: “Describe the olfactory characteristics of the wine”
- 140 - Gustatory: “Describe the gustatory characteristics of the wine”

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

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

145 *2.1.1.4. CATA task and datasets*

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

150 “Check in the subsequent list the words that apply to this wine”.

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

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

166 **2.1.2. Second study: milk chocolates**

167 *2.1.2.1. Participants*

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

176 *2.1.2.2. Products*

177 Four milk chocolate with different recipes were used: a standard Belgian milk
178 chocolate, a Swiss milk chocolate, a milk compound chocolate, and a protein base milk
179 chocolate.

180 *2.1.2.3. FC task and datasets*

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

- 184 - Mouth texture: “Describe the mouth texture characteristics of the chocolate”
- 185 - Mouth flavor: “Describe the mouth flavor characteristics of the chocolate”

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

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

190 *2.1.2.4. CATA task and datasets*

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

194 “Check in the subsequent list the words that apply to this chocolate”.

195 The CATA lists of mouth texture and mouth flavor descriptors were composed of 8 and
196 6 descriptors respectively. The mouth texture descriptors were the following: hard, soft,
197 sticky, melting, coarse, fatty, creamy texture, and mouthcoating. The mouth flavor
198 descriptors were the following: sweet, bitter, cocoa, caramel, cereal, and milky. These
199 descriptors were selected according to the expertise of Barry Callebaut© and were
200 presented in a different randomized order for each consumer but with a constant order
201 across evaluations for a given consumer.

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

205 *2.2. Data treatment*

206 *2.2.1. FC data treatment*

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

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

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

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

220 2.2.2. CATA data treatment

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

225 2.3. Data analyses

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

227 2.3.1. Similarity of FC and CATA outputs

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

231 2.3.2. Size of the differences between the products

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

234

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

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

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

255 **2.3.3. Stability of the outputs**

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

260 The stability of the descriptor configurations was not investigated ([Ares, Bruzzone, et
261 al., 2014](#); [Ares, Tárrega, et al., 2014](#); [Vidal et al., 2015](#)) because they are usually not
262 interpreted for themselves but rather as help for interpretation to understand the
263 product configurations. In this sense, the stability of the joint product by descriptor
264 configurations and of the product by descriptor significant associations were

265 investigated instead. The choice to keep two indicators (joint product by descriptor
266 configurations and product by descriptor significant associations) that seem similar is
267 deliberate. The joint product by descriptor configurations corresponds to the product
268 by descriptor insights one would draw from reading the map and/or the space resulting
269 from the CA of the contingency table. By nature, this reading is subjective and
270 approximate but has the benefit of being nuanced. The product by descriptor significant
271 associations are the black and white version of the joint product by descriptor
272 configurations and corresponds to the product by descriptor insights one would draw
273 from reading the tables as presented [Mahieu et al. \(2020\)](#). By their statistical-based
274 nature, the product by descriptor significant associations are objective but have the
275 drawback of being threshold-dependent and binary.

276 *2.3.3.1. Bootstrap resampling procedure*

277 For each dataset, different sizes of virtual panels were considered ranging from 10 to
278 the size of the actual panel, increasing with a step of 10. For each size, 1000 virtual
279 panels were constituted. Each virtual panel was constituted by randomly drawing
280 subjects from the actual panel with replacement. The outputs obtained from the actual
281 panel were considered [as a benchmark to which the outputs of the virtual panels were](#)
282 [compared](#).

283 *2.3.3.2. Product configurations*

284 The product configurations, i.e. the relative position of the products in relation to each
285 other in the sensory space, were compared by computing the RV coefficient ([Escoufier,](#)
286 [1973; Robert & Escoufier, 1976](#)) in the full space between the product configurations
287 of the actual and the virtual panels.

288 *2.3.3.3. Joint product by descriptor configurations*

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

295 *2.3.3.4. Product by descriptor significant associations*

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

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

$$303 \quad \phi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

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

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

314 *2.3.3.5. Stability of outcomes*

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

326 To compare the 5% quantile of the distributions of the correlation coefficients to the
 327 different thresholds rather than the mean of these distributions (Ares, Bruzzone, et al.,
 328 2014; Ares, Tárrega, et al., 2014; Blancher et al., 2012; Cadena et al., 2014; Vidal et
 329 al., 2014) is more in line to what a virtual panel drawn from the bootstrap resampling
 330 of the actual panel represents. Indeed, under the hypothesis where such a virtual panel
 331 represents a new study conducted in similar experimental settings, similar outputs to
 332 those of the actual panel considered as a benchmark are expected from this virtual
 333 panel. Thus, high correlation coefficients between the outputs of the actual and the
 334 virtual panel are expected. Extended to a large number of virtual panels, this line of
 335 reasoning still holds, and thus considering the entire distribution rather than its mean
 336 is more in line with the bootstrap hypothesis made and with what a virtual panel
 337 represents.

338 3. Results

339 3.1. Similarity of FC and CATA outputs

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

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

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

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

348 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.

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

359 [3.3. Stability of the outputs](#)

360 [3.3.1. Product configurations](#)

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

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

374 [3.3.2. Joint product by descriptor configurations](#)

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

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

391 3.3.3. Product by descriptor significant associations

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

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

405 4. Discussion

406 4.1. The stability of the outputs provided by FC and CATA

407 Results showed relatively stable FC outputs, at least as stable as CATA ones if not
408 more. FC outputs reached good stability in more datasets than CATA ones regarding
409 product configurations and joint product by descriptor configurations. Further, the
410 average stability of FC outputs was always larger than or equal to CATA ones for the

411 three aspects of the outputs investigated in this study when a given pair product /
412 sensory-modality with 60 consumers and more was considered. [These results suggest](#)
413 [that FC outputs are on the same level of stability that CATA ones, at least when FC](#)
414 [and CATA are performed by sensory modality](#). Future studies need to be conducted to
415 confirm or refute these results when FC and CATA are performed with a single overall
416 characterization of each product (not by sensory modality).

417 [The previous statements worth being nuanced by two points. First, the consumers who](#)
418 [performed the chocolate CATA task might be more knowledgeable about chocolate](#)
419 [than if they were naïve consumers. Thus, the CATA descriptions might have been](#)
420 [more consensual, which might have resulted in higher stability of the outputs.](#)
421 [Therefore, the stability of CATA outputs might have been overestimated in the](#)
422 [chocolate study. Second, some descriptors of the CATA list in the wine study may be](#)
423 [considered reasonably technical \(e.g. animal, roasted, etc.\). This may have impeded](#)
424 [the agreement of consumers on CATA descriptions, which may have resulted in lesser](#)
425 [stability of the outputs. However, some of these “technical descriptors” were mentioned](#)
426 [during the FC task \(Mahieu et al., 2020\), which suggests that they were meaningful to](#)
427 [consumers. They were however mentioned less frequently in FC as compared to](#)
428 [CATA, but so were common descriptors shared by FC and CATA \(Mahieu et al., 2020\).](#)
429 [Indeed, the CATA task encourages consumers to check the proposed descriptors](#)
430 [\(Callegaro, Murakami, Tepman, & Henderson, 2015; Kim, Hopkinson, van Hout, & Lee,](#)
431 [2017; Krosnick, 1999\). This suggests that this difference in citation frequency is due to](#)
432 [the task and not to the potential “technical” aspect of the descriptors.](#)

433 Not surprisingly, for both FC and CATA, the stability of the product configurations
434 increased with the size of the virtual panel and with the size of the differences between
435 the products. The minimum number of subjects to obtain stable product configurations
436 was of the same order of magnitude that was previously reported for CATA, RATA,
437 Projective Mapping, Sorting, and Polarized Sensory Positioning ([Ares, Bruzzone, et](#)
438 [al., 2014; Ares, Tárrega, et al., 2014; Blancher et al., 2012; Cadena et al., 2014; Vidal](#)
439 [et al., 2015](#)).

440 The overall level of stability was more impacted by the size of product differences than
441 by the method used (FC versus CATA). These results are in line with some previously
442 reported studies ([Ares, Bruzzone, et al., 2014; Ares, Tárrega, et al., 2014; Blancher et](#)
443 [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.

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

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

491 5. Conclusion

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

506 Acknowledgments

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

509 The authors would like to thank Robert et Marcel®, Sicarex®, and Barry Callebaut®
510 for providing their products.

511 References

- 512 Adams, J., Williams, A., Lancaster, B., & Foley, M. (2007). Advantages and uses of
513 check-all-that-apply response compared to traditional scaling of attributes for
514 salty snacks. In, *7th Pangborn Sensory Science Symposium*. Minneapolis,
515 USA.
- 516 Ares, G., Bruzzone, F., Vidal, L., Cadena, R. S., Giménez, A., Pineau, B., et al. (2014).
517 Evaluation of a rating-based variant of check-all-that-apply questions: Rate-all-
518 that-apply (RATA). *Food Quality and Preference*, *36*, 87-95.
- 519 Ares, G., Giménez, A., Barreiro, C., & Gámbaro, A. (2010). Use of an open-ended
520 question to identify drivers of liking of milk desserts. Comparison with
521 preference mapping techniques. *Food Quality and Preference*, *21*(3), 286-294.
- 522 Ares, G., Tárrega, A., Izquierdo, L., & Jaeger, S. R. (2014). Investigation of the number
523 of consumers necessary to obtain stable sample and descriptor configurations
524 from check-all-that-apply (CATA) questions. *Food Quality and Preference*, *31*,
525 135-141.
- 526 Blancher, G., Clavier, B., Egoroff, C., Duineveld, K., & Parcon, J. (2012). A method to
527 investigate the stability of a sorting map. *Food Quality and Preference*, *23*(1),
528 36-43.
- 529 Cadena, R. S., Caimi, D., Jaunarena, I., Lorenzo, I., Vidal, L., Ares, G., et al. (2014).
530 Comparison of rapid sensory characterization methodologies for the
531 development of functional yogurts. *Food Research International*, *64*, 446-455.
- 532 Callegaro, M., Murakami, M. H., Tepman, Z., & Henderson, V. (2015). Yes-no answers
533 versus check-all in self-administered modes. *International Journal of Market*
534 *Research*, *57*, 203-223.
- 535 Castura, J. C., Antúnez, L., Giménez, A., & Ares, G. (2016). Temporal Check-All-That-
536 Apply (TCATA): A novel dynamic method for characterizing products. *Food*
537 *Quality and Preference*, *47*, 79-90.
- 538 Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.).
539 Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- 540 Cramér, H. (1946). Chapter 21. The two-dimensional case. In P. U. Press,
541 *Mathematical Methods of Statistics*.
- 542 Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, *29*, 751-760.
- 543 Fleming, E. E., Ziegler, G. R., & Hayes, J. E. (2015). Check-all-that-apply (CATA),
544 sorting, and polarized sensory positioning (PSP) with astringent stimuli. *Food*
545 *Quality and Preference*, *45*, 41-49.
- 546 Gacula Jr, M., & Rutenbeck, S. (2006). Sample size in consumer test and descriptive
547 analysis. *Journal of Sensory Studies*, *21*(2), 129-145.
- 548 Greenacre, M. (2013). Contribution Biplots. *Journal of Computational and Graphical*
549 *Statistics*, *22*(1), 107-122.
- 550 Hanaei, F., Cuvelier, G., & Sieffermann, J. M. (2015). Consumer texture descriptions
551 of a set of processed cheese. *Food Quality and Preference*, *40*, 316-325.

552 Heymann, H., Machado, B., Torri, L., & Robinson, A. L. (2012). How many judges
553 should one use for sensory descriptive analysis? *Journal of Sensory Studies*,
554 27(2), 111-122.

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

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

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

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

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

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

573 Mammasse, N., & Schlich, P. (2014). Adequate number of consumers in a liking test.
574 Insights from resampling in seven studies. *Food Quality and Preference*, 31,
575 124-128.

576 Oppermann, A. K. L., de Graaf, C., Scholten, E., Stieger, M., & Piqueras-Fiszman, B.
577 (2017). Comparison of Rate-All-That-Apply (RATA) and Descriptive sensory
578 Analysis (DA) of model double emulsions with subtle perceptual differences.
579 *Food Quality and Preference*, 56, 55-68.

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

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

584 Reinbach, H. C., Giacalone, D., Ribeiro, L. M., Bredie, W. L. P., & Frøst, M. B. (2014).
585 Comparison of three sensory profiling methods based on consumer perception:
586 CATA, CATA with intensity and Napping®. *Food Quality and Preference*, 32,
587 160-166.

588 Robert, P., & Escoufier, Y. (1976). A Unifying Tool for Linear Multivariate Statistical
589 Methods: The RV- Coefficient. *Applied Statistics*, 25(3), 257-265.

590 Silva, R. d. C. d. S. N. d., Minim, V. P. R., Silva, A. N. d., & Minim, L. A. (2014). Number
591 of judges necessary for descriptive sensory tests. *Food Quality and Preference*,
592 31, 22-27.

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

596 ten Kleij, F., & Musters, P. A. D. (2003). Text analysis of open-ended survey responses:
597 a complementary method to preference mapping. *Food Quality and Preference*,
598 14(1), 43-52.

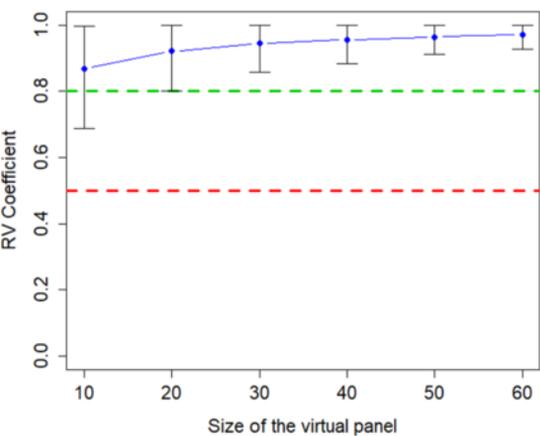
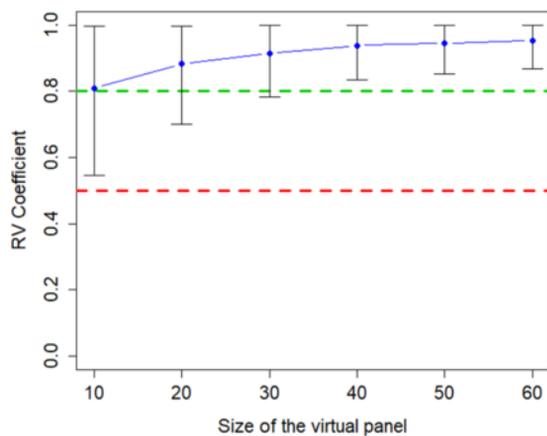
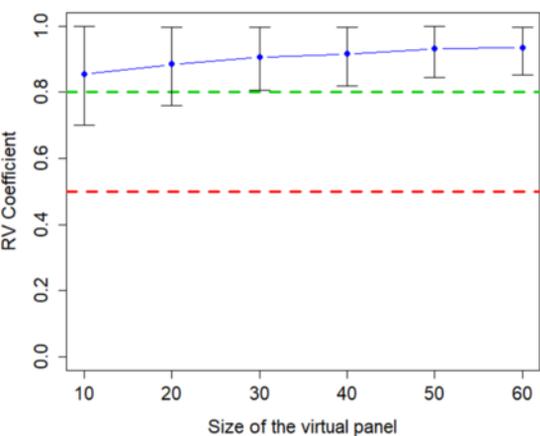
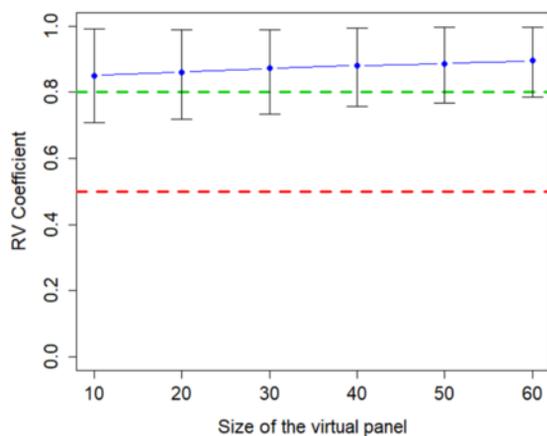
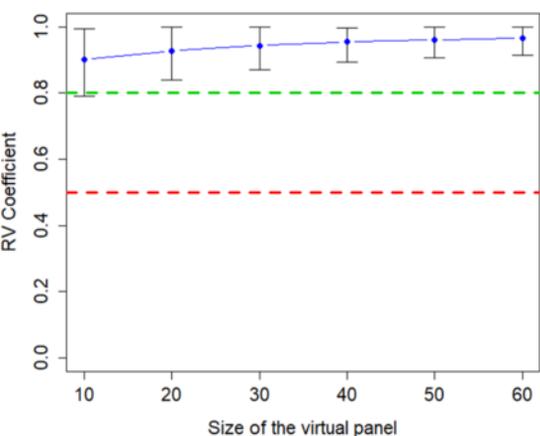
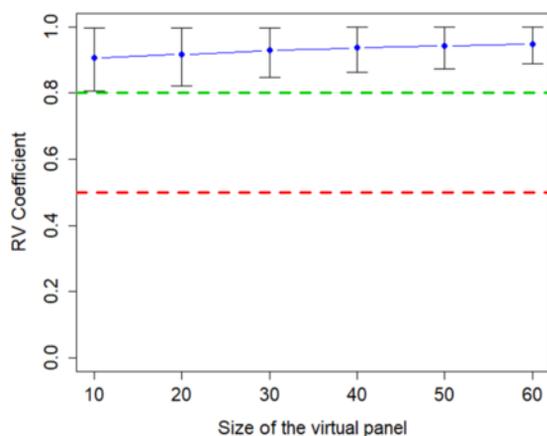
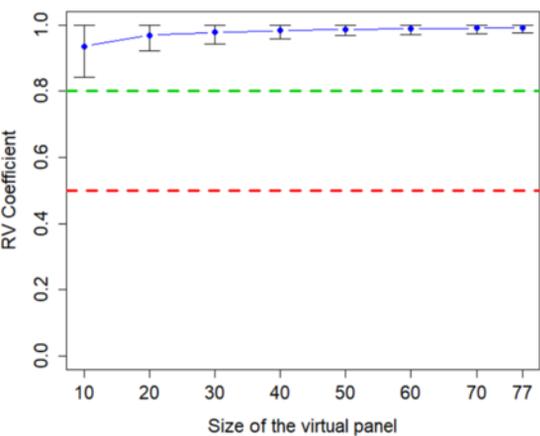
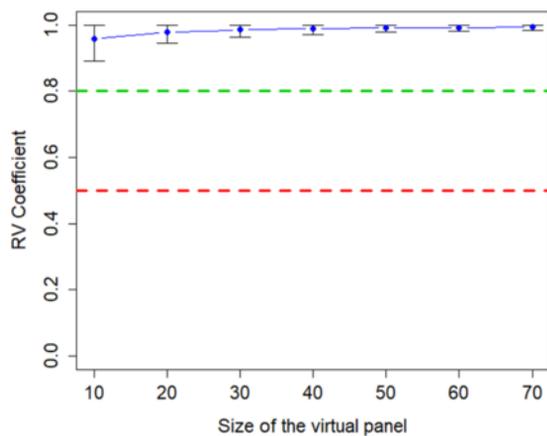
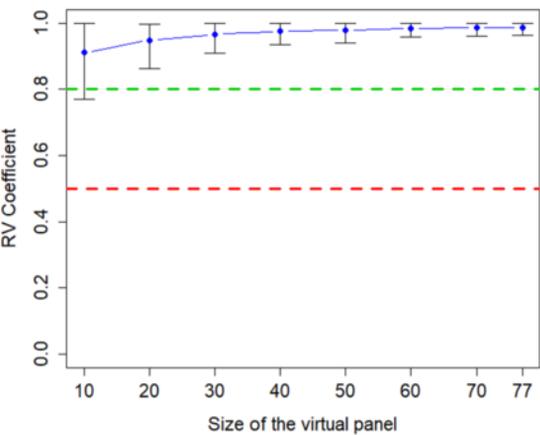
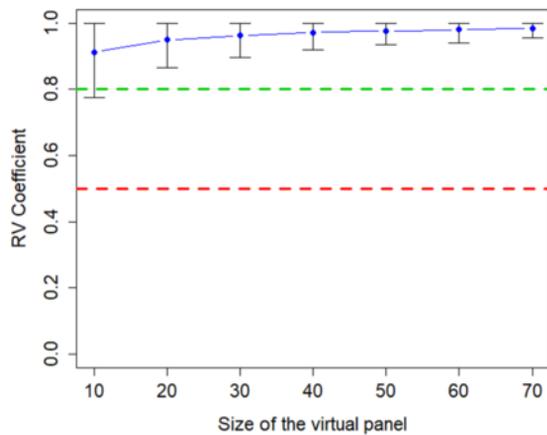
599 Valentin, D., Chollet, S., Lelièvre, M., & Abdi, H. (2012). Quick and dirty but still pretty
600 good: a review of new descriptive methods in food science. *International Journal*
601 *of Food Science & Technology*, 47(8), 1563-1578.

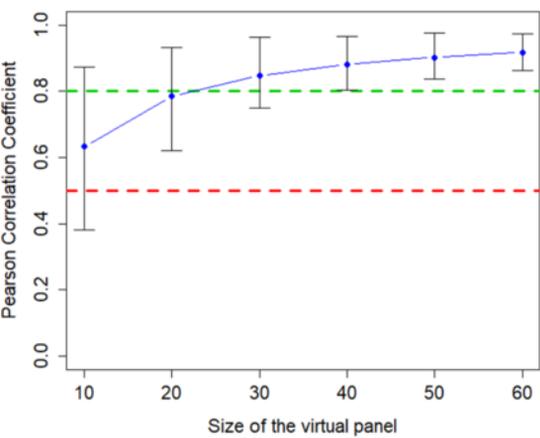
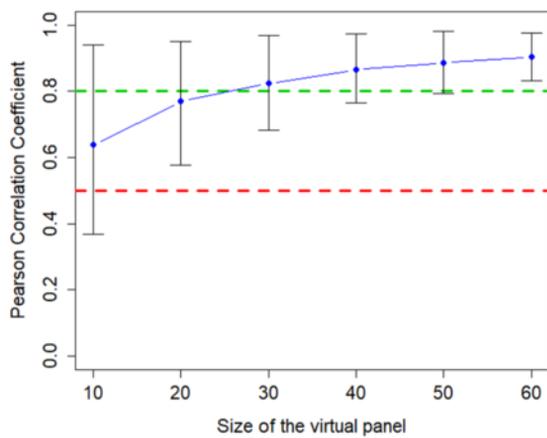
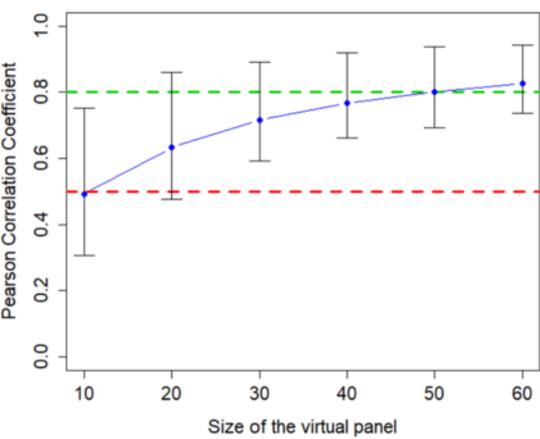
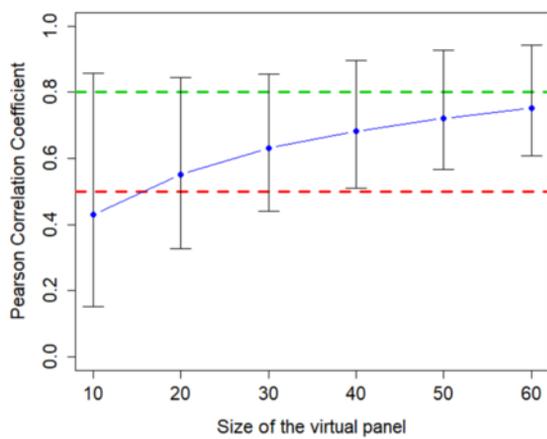
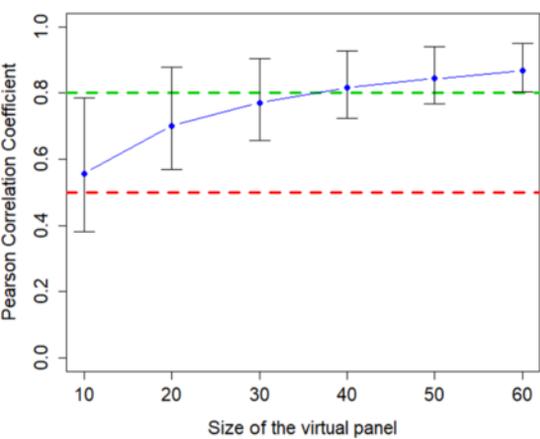
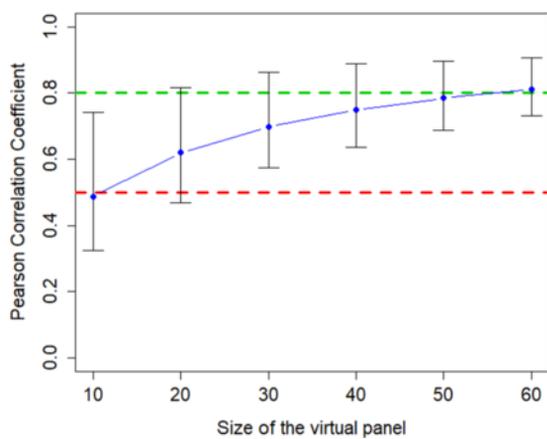
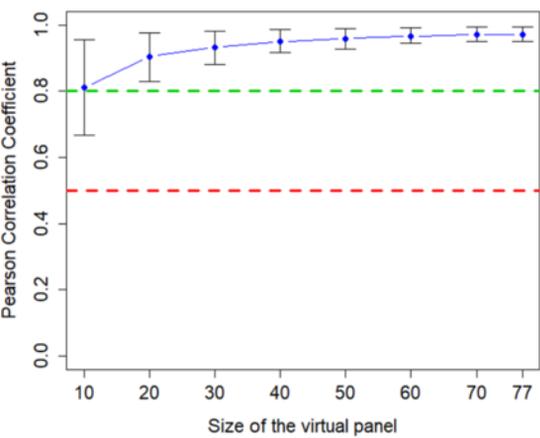
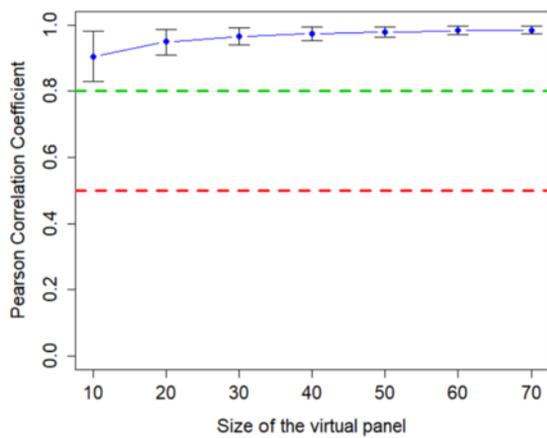
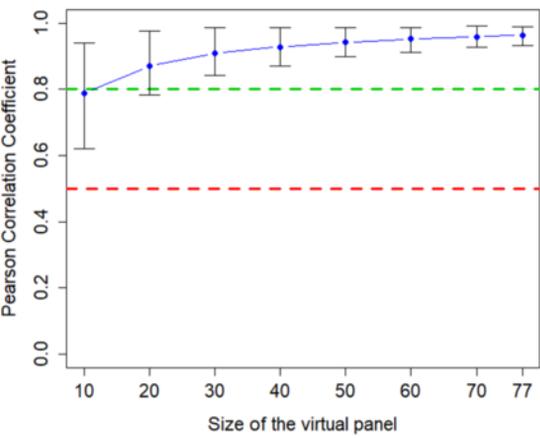
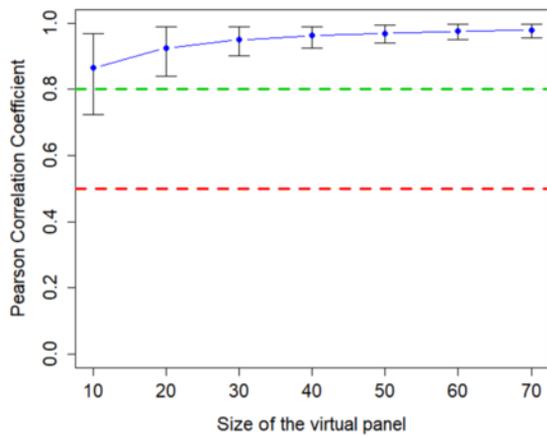
- 602 Varela, P., & Ares, G. (2012). Sensory profiling, the blurred line between sensory and
603 consumer science. A review of novel methods for product characterization.
604 *Food Research International*, 48(2), 893-908.
- 605 Vidal, L., Cadena, R. S., Antúnez, L., Giménez, A., Varela, P., & Ares, G. (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.

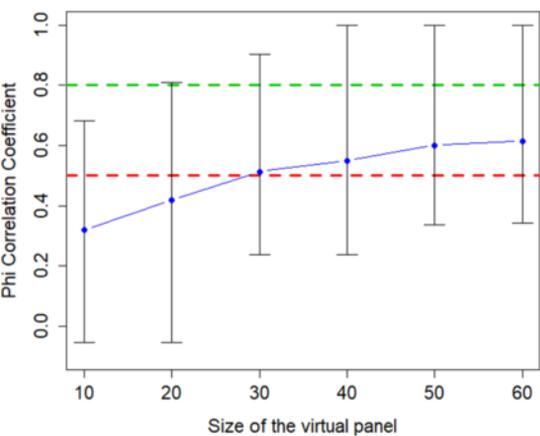
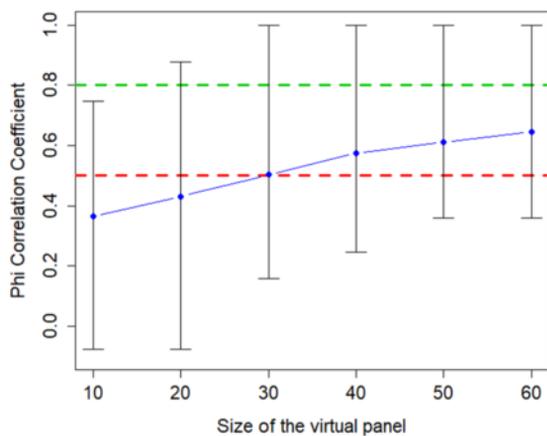
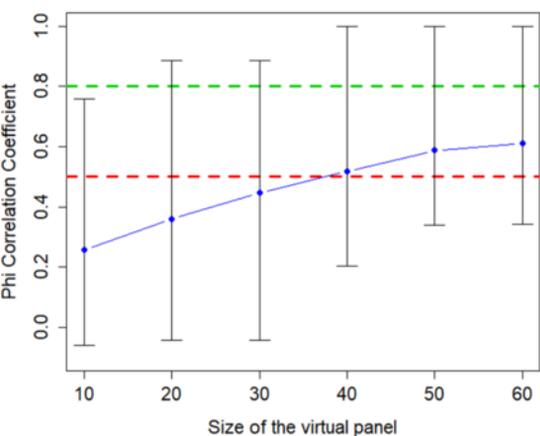
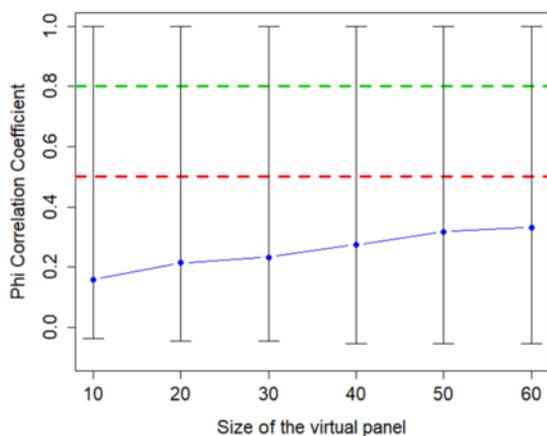
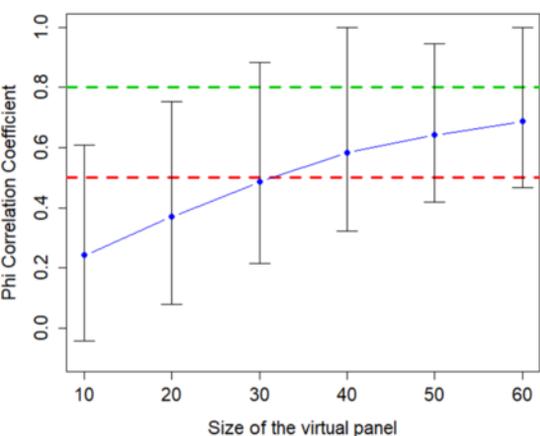
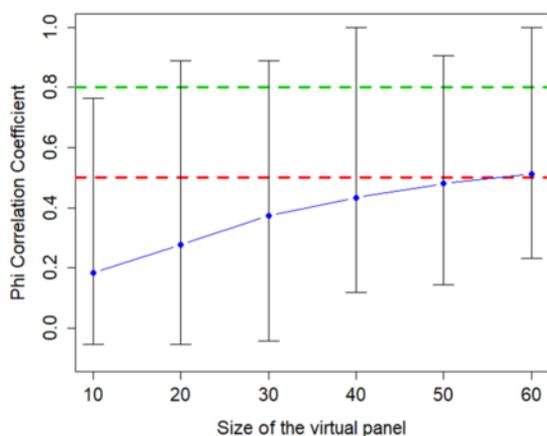
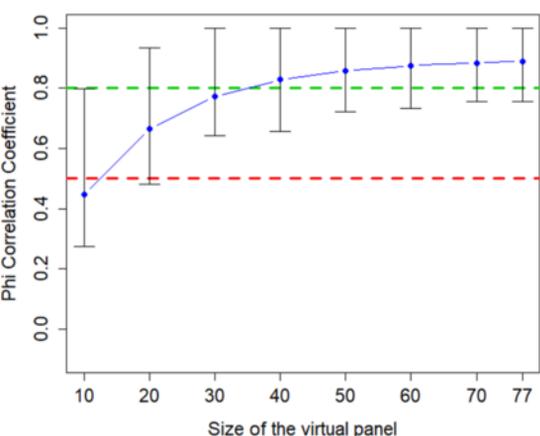
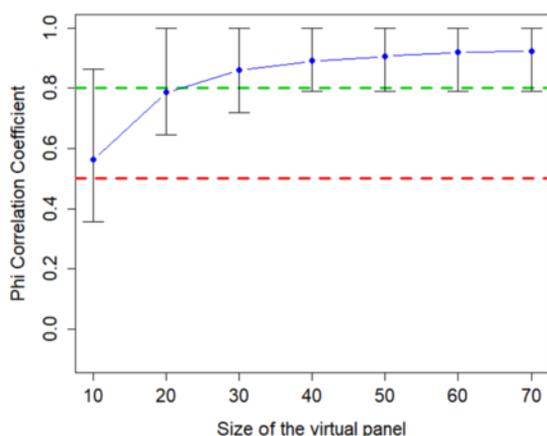
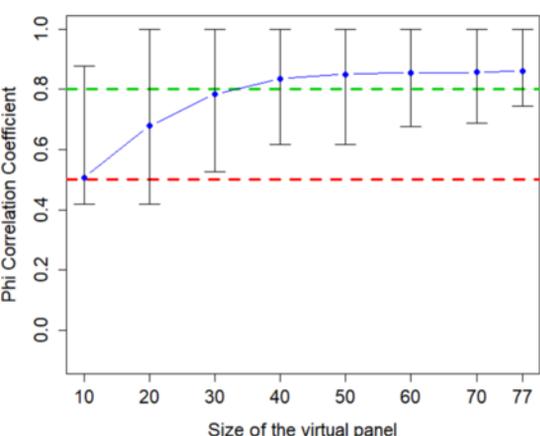
Fig. 1: Mean of the distribution of the RV coefficients between the actual and the virtual product configurations as a function of the virtual panel size for (a) FC-Wine-Vis, (b) CATA-Wine-Vis, (c) FC-Wine-Olf, (d) CATA-Wine-Olf, (e) FC-Wine-Gus, (f) CATA-Wine-Gus, (g) FC-Choc-Tex, (h) CATA-Choc-Tex, (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 ($\alpha = 5\%$) 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)**

(a)**(b)****(c)****(d)****(e)****(f)****(g)****(h)****(i)****(j)**

(a)**(b)****(c)****(d)****(e)****(f)****(g)****(h)****(i)****(j)**