Food Quality and Preference 51 (2016) 65–71

Contents lists available at ScienceDirect

Food Quality and Preference

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

Short Communication

Rate-all-that-apply (RATA) with semi-trained assessors: An investigation of the method reproducibility at assessor-, attribute- and panel-level

Davide Giacalone [a,*], Pia Ingholt Hedelund [b]

[a] Department of Technology and Innovation, University of Southern Denmark, Odense, Denmark
[b] Danish Technological Institute, Århus, Denmark

ARTICLE INFO

Article history:
Received 12 October 2015
Received in revised form 18 February 2016
Accepted 24 February 2016
Available online 24 February 2016

Keywords:
Rate-all-that-apply
RATA
Assessor reproducibility
Panel agreement
Attribute stability
Sensory profiling
Chocolate

ABSTRACT

Rate-all-that-apply (RATA) is a variant of check-all-that-apply (CATA) questions that allows assessor to rate the intensity of selected attributes. Compared to CATA, RATA has the potential to improve sample description and discrimination, and might be more useful when only a small number of assessors are available. Before advocating its use with confidence, investigations on the method validity and reproducibility are necessary.

Within this context, this short paper examined the reproducibility of results obtained by RATA within a test–retest paradigm, drawing on data from a case study involving sensory assessment of common defects in chocolate production. Criteria considered were within-assessors reproducibility, attribute stability, and configurational agreement between samples spaces obtained across replicated evaluations. The results showed that although within-assessors reproducibility was moderate, RATA exhibited a very good reproducibility at panel level, as indicated by the high configurational agreement between product maps obtained from individual replicates. The method showed a good reproducibility also at the level of individual attributes. Indications were obtained that the reproducibility of RATA with semi-trained subjects might be similar to that of a simple checklist, in spite of the addition of the intensity rating step.

Overall, the work presented in this short paper supports the validity of RATA as a sensory profiling tool, and suggests that its application with semi-trained assessors may be particularly advantageous for industrial applications.

1. Introduction

1.1. Rate-all-that-apply (RATA)

Check-all-that-apply (CATA) questions are a fast product profiling technique consisting in presenting assessors with a product and checklist of predefined attributes, from which the assessor is asked to select the ones he or she finds appropriate for describing the sample (Adams, Williams, Lancaster, & Foley, 2007). This method has shown several advantages in terms of reproducibility, ease of use, and rapidity which have contributed to its increasing popularity (Meyners & Castura, 2014). The lack of bias on hedonic response (with regards to concurrent elicitation of sensory and liking on the same ballot) has also contributed to its widespread adoption, particularly with regards to its application with consumer panels (Jaeger et al., 2013b).

On the other hand, CATA questions present some limitations that may limit their use in other contexts than large-scale consumer studies. The most relevant ones are that (1) CATA produces dichotomous data (1/0) which may lack sufficient power to discriminate between samples with relatively subtle sensory differences (Reinbach, Giacalone, Ribeiro, Bredie, & Frøst, 2014), thus requiring a substantially large sample size (Ares, Tárrega, Izquierdo, & Jaeger, 2014c), and (2) that the method does not encourage a deep processing of the attributes on the ballot and may therefore prompt "satisficing" strategies in the assessors (Meyners & Castura, 2014).

To obviate these shortcomings, some authors have proposed a rating-based variant of CATA (Reinbach et al., 2014), in which assessors are required to evaluate the intensity of every applicable attribute, an approach referred to as "rate-all-that-apply" or RATA (Ares et al., 2014b). At an overall level, comparisons of RATA and CATA have shown that the two methods provide similar information (Ares et al., 2014c; Reinbach et al., 2014), but also some evidence for a greater discriminative capacity of the RATA format (Ares et al., 2014c).

[*] Corresponding author at: Campusvej 55, DK-5230 Odense M, Denmark.
E-mail address: dg@iti.sdu.dk (D. Giacalone).

http://dx.doi.org/10.1016/j.foodqual.2016.02.017
0950-3293/© 2016 Elsevier Ltd. All rights reserved.

To date, RATA has received some attention mostly with regards to comparisons with its CATA counterpart in applications with consumers (Ares et al., 2014b; Jaeger & Ares, 2015; Reinbach et al., 2014). However, little is known about other methodological aspects of RATA such as its method reproducibility. To address this paucity of information, this paper seeks to investigate the reproducibility of the results obtained across replicated evaluations using the RATA method.

1.2. Aims of the present research

The larger context of this research was a collaboration with a confectionary company, which was interested in using a fast sensory method in their quality control (QC) of their chocolate production. Several comprehensive approaches for the use of sensory methods in QC have been proposed (Civille, Carr, & Munõz, 1992), but they generally require substantial resources, and the necessary number of qualified employees is not always available. Quicker methods that can be employed with a few assessors, possibly using the company own employees, would therefore be advantageous (Costell, 2002).

Among the several fast sensory methods proposed in the literature (see e.g., Varela & Ares, 2012), the RATA method has the potential to deliver on this need. There are several reasons why this is the case. First, since the method is based on a pre-defined list of attributes, the ballot can be tailored to describing key production errors, which is important in QC applications (Costell, 2002). Moreover, in addition to the advantages it shares with the CATA format (speed, cognitive ease), RATA is known to increase processing of the ballot and the number of attributes used by the assessors (Ares et al., 2014b). Finally, the rating step increases its discriminating power which is important when working with small panels, and/or when sensory differences between the test samples are subtle. The latter is often the case in QC where a defect sample might share several attributes with the reference product, but differ with regards to their intensity.

However, before advocating its adoption with confidence, it is important to ascertain whether this method produces valid results.

In this research, we focus on reliability, i.e. the degree to which a method produces results that are stable and reproducible across repeated evaluations. Reliability is a relevant facet of validity to focus on in the present context, because in daily QC applications replication may not be an option due to practical constraints.

Method reliability can be investigated within a test–retest paradigm, that is, by obtaining responses from the same group of assessors at different time points, such as when replicate evaluations are conducted on the same sample set. This has been done in similar research on CATA questions, both with trained assessors (Campo, Do, Ferreira, & Valentin, 2008) and consumers (Jaeger et al., 2013; Ares et al., 2014a).

This work extends such methodological investigation to the RATA approach. Specifically, the purpose of this paper is to investigate the reproducibility of the results from a RATA task collected using a semi-trained panel performing replicated evaluations. Three main evaluative criteria will be addressed:

(1) assessor reproducibility: the degree to which individual (semi-trained) assessors use a RATA ballot similarly across replicate evaluations;
(2) attribute stability: at the panel level, the degree to which individual attributes are used identically across replicated evaluations;
(3) reproducibility of global sensory characterizations: at the panel level, the degree to which sample spaces obtained in replicated evaluations are congruent, and whether they lead to (dis)similar sensory conclusions about the samples.

2. Materials and methods

2.1. Samples

A total of eight chocolate samples were produced in a small production site and represented standard recipes (i.e., chocolate without production errors), and a range of critical production errors, well known in the chocolate industry, that are not easily detectable by instrumental or microbiological methods. Table 1 provides a list and brief description of the samples.

Sample B provided the basis for all the chocolate defects (samples D–H) which represented production errors relative to this standard. Samples A and C are standard recipes and were introduced to increase the complexity of the task for the assessors and to cover a wider and more multidimensional product space. As an additional diagnostics measure, sample B was also used as a blind duplicate in all sensory evaluations, meaning that the assessors evaluated nine chocolate in total. Samples are referred to by the product codes indicated in Table 1 in the remainder of the paper.

2.2. Panelists

An in-house panel composed of 16 company employees was initially recruited on the basis of interest and availability. All assessors had considerable technical expertise about chocolate production, but little or no prior experience with formal sensory evaluation.

The panel underwent four training sessions (35–45 min each), during which the assessors were introduced to sensory science, received instructions and reference materials for the RATA ballot (see Section 2.3), and were exposed to the focal samples. Additionally, they were screened for sensory acuity in relation to the basic tastes. The screening procedure was conducted in accordance with ISO 8586-1 (1993) and included both a recognition test and a threshold test. Five assessors were excluded from the panel during the training phase: three of them because they could not complete the training, and two of them because they performed below expectations during the screening procedure.[1]

The final panel was therefore composed of 11 assessors (7 women, 21–51 years of age).

2.3. Vocabulary development and RATA ballot

A trained external panel ($N = 6$) from the University of Copenhagen, Denmark was used to develop an initial list of attributes. To this end, the trained assessors evaluated all the samples in four consecutive sessions and wrote down all attributes they could think of for describing the samples considering all relevant sensory modalities. The sensory attributes most frequently mentioned, and for which a reference or a common definition could be identified, were chosen for inclusion in the RATA ballot. The final list comprised 65 attributes. For brevity, these are not discussed but interested readers will find them reported in Table 3.

The RATA ballot listed the attributes broken down by sensory modalities, because this format of presentation has been reported to improve attribute processing and reduce cognitive burden in similar tasks (Ares & Jaeger, 2013). Within sensory modalities, attributes appeared in a fixed order. To further ease attribute processing, the order in which the modalities appeared was in line with the expected 'dynamics of sensory perception' (Ares &

[1] At the request of the company, even those two assessors who did not meet the minimum performance according to ISO 8586-1 were allowed to take part in the study. However, their results were omitted from the analysis.

Table 1
Description of the chocolate samples used in the study. Samples "B-1" and "B-2" came from the same production batch and were used as blind duplicates. The same sample is also the standard to which the defects (samples D–H) relate.

Product codes	Cocoa content (%)	Description
A	57	Standard recipe
B-1	70	Standard recipe
B-2	70	Standard recipe
C	80	Standard recipe
D	70	Defect sample obtained by storing the chocolate under temperature fluctuation, resulting in so-called "bloomed" chocolate (Briones & Aguilera, 2005)
E	70	Defect sample obtained by addition of extra lecithin (4 g per 100 g chocolate mass)
F	70	Defect sample obtained by overly roasting the cocoa nibs
G	70	Defect sample obtained by skipping the "conching" process (Beckett, 2009)
H	70	Defect sample obtained by prolonged conching (twice the time as normal production for sample B)

Jaeger, 2013; Ares et al., 2013): (1) appearance, (2) odor, (3) flavor/taste, (4) texture and (5) aftertaste.

The task for the assessors consisted in ticking all the attributes they perceived in the samples and evaluate their intensity on a 3-point scale with all points labeled (1 = "Low", 2 = "Medium", 3 = "High"). This RATA format was identical to the one used in Study 2 and 3 of Ares et al. (2014b).

2.4. Test procedures

The sensory evaluation of the samples was carried out at the company's own facilities, which included a room for preparing the samples, a training room, and 10 individual booths.

During the sensory evaluations, samples (a single piece of 10 g) were served blinded in a transparent polystyrene cup with a lid, labeled with a three digit code. The samples were stored at 15 °C for 24 h and brought to room temperature (20 °C) prior to serving. The samples were presented monadically in a randomized serving order according to a complete block design. Cucumber, lukewarm water and cold water were used as palate cleansers in between samples. The assessors were requested not to eat, drink or smoke at least one hour prior to the tests.

Sensory evaluations using the RATA method was conducted in four replicates over a month period, with each session lasting approximately 40 min.

2.5. Data analysis

A mix of univariate and multivariate statistical analyses were performed in order to address the aims outlined in the introduction. All analyses were performed using R version 3.0.2 (R Development Core Team, 2013), using either native functions or functions from the FactoMineR package (Lê, Josse, & Husson, 2008).

2.5.1. Quantitative indices for assessors and attribute reproducibility

For ordinary CATA questions, two quantitative indices for assessment of within-assessors reproducibility have recently been recently proposed. The first one is the "average reproducibility index" by Campo et al. (2008), defined as follows:

$$R_i = \frac{r}{n} \times \sum_{j=1}^{n} \left(\frac{des_{com\,j}}{des_{rep\,(1)j} + des_{rep\,(2)j} + \cdots + des_{rep\,(r)j}} \right) \tag{1}$$

where n is the number of samples, r is the number of replicates, $des_{com\,j}$ is the number of common attributes used by assessor i in the r replicates for describing sample j, $des_{rep\,(1)}, \ldots, des_{rep\,(r)}$ are the number of attributes used by assessor i in each replicate, and R_i is the average reproducibility for assessor i, ranging from 0 to 1 (1 = perfect reproducibility).

The Campo et al. index was proposed for a specific protocol that limited the number of attributes assessors could choose from the full list. If applied directly to ordinary CATA questionnaire, then this index has the drawback that does not account for items that are left unchecked in CATA questionnaires. Thus, Jaeger et al. (2013a) proposed an alternative index that takes into account the whole set of CATA items:

$$RI_i = \frac{1}{n} \times \sum_{j=1}^{n} \left(\frac{des_{com\,j}}{des} \right) \tag{2}$$

where $des_{com\,j}$ is the number of items used by assessor i identically in all replicates for sample j, des is the total number of items present in the ballot, n is the number of samples, and RI_i is the global reproducibility index for assessor i, again ranging from 0 to 1, where 1 indicates perfect reproducibility.

Both the index proposed by Campo et al. (2008) and the one proposed by Jaeger et al. (2013a) are readily applicable to the present experimental design, as they quantify assessor reproducibility across replicates. However, considering the protocol used for the evaluation, taking into account unchecked items was considered important, and thus the global reproducibility index RI_i was considered as a primary measure of assessor reproducibility in this study.

In a recent paper, Worch and Piqueras-Fiszman (2015) observed that the RI_i index might be too optimistic if the experimental conditions inflate the amount of unchecked items. This could be the case, for example, if assessors are restricted to select a limited number of attributes from a large list (as in Campo et al., 2008), if there are attributes clearly not applicable to the sample space, or if long lists of attributes are used with consumer panels who are adopting satisficing response strategies (Meyners & Castura, 2014). Nevertheless, we considered the index proposed by Jaeger et al. (2013a) to be appropriate in the present context, considering that the protocol of this study left assessors free to choose any number of attributes. Critically, the fact that semi-trained assessors (experienced with the ballot and trained on the specific attributes) were employed for the evaluation makes it more likely that unchecked items were intentionally left blank.

In addition to assessor reproducibility, quantitative indices for reproducibility at the level of individual attributes have also been proposed. For CATA data, Jaeger et al. (2013a) developed an "attribute stability index" to assess the percentage of assessors that use each attribute identically across samples and replicates, which we used in this paper to quantify reproducibility at the attribute level. It is calculated as follows (Jaeger et al., 2013a):

$$SI_k = \frac{100}{n \times N} \times \sum_{i=1}^{N} (sam_i) \tag{3}$$

where sam_i is the number of samples for which assessor i used the attribute k identically across all replicates, n is the number of samples, N is the number of assessors, and SI_k is the stability index of attribute k, ranging from 0 to 100 (100 = perfect stability). It should be noted here that both the RI_i and SI_k indices have previously been used to quantify reproducibility across two replicates only, whereas in the present work we extend their usage to a higher number replicates.

Additionally, these indices were initially developed for evaluating simple frequency data. Since it was interesting to assess the reproducibility of RATA also vis-à-vis simple frequencies, we decided to use these indices in two ways in the context of this work. First, we dichotomized the data into a CATA-like form, and considered $des_{com\,j} = 1$ and $sam_i = 1$ satisfied if the attributes was used identically across all replicates simply with regards to their being checked or not. This is the typical situation when only the frequency of mention of an attribute is considered, therefore we refer to these variants as RI_i^{FREQ} and SI_k^{FREQ}.

Then, we considered the harsher criterion $des_{com\,j} = 1$ and $sam_i = 1$ if attributes were used identically across replicates with regards to the actual scale point. These variants are referred to RI_i^{RATA} and SI_k^{RATA}.

2.5.2. Reproducibility of sensory characterizations

The third aim of the paper was to assess reproducibility of the whole sensory characterization. This can be achieved, through ad hoc multivariate data analyses, by evaluating the configurational similarity of product spaces obtained from separate replicates. In this paper, we used Multiple Factor Analysis (MFA, Escoufier & Pagès, 1994). MFA is a method which aims at integrating different groups of variables describing the same samples, defining summed data from individual replicates as the initial groups of variables. The reproducibility of the product spaces was then evaluated through the typical MFA visuals (sample space and partial point representation). The interpretation was also aided by the presence of a blind duplicate (Table 1).

Additionally, the RV coefficient (Robert & Escoufier, 1976) was used to measure similarities between MFA sample configurations obtained from different replicates. The RV coefficient is a multivariate generalization of the simple correlation coefficient and takes value between 0 and 1, where 1 indicates perfect configurational congruence between the two matrices.

In the present context, the RV coefficient was computed for all possible replicate combinations considering two (the standard number of dimensions of interest in most applications), four, and all MFA dimensions.

3. Results and discussion

3.1. Assessor reproducibility

The first evaluative criterion considered in this paper was within-assessor reproducibility. Assessors used on average 15.6 attributes (24% of the total) per sample. Table 2 presents reproducibility indices for the 11 assessors with the Global Reproducibility Index RI_i. Assessors' global reproducibility index (RI_i^{RATA}) spanned a range between 0.66 (Assessor 1) and 0.45 (Assessor 7). The mean value was 0.55, indicating that on average 55% of the terms (36 out of 65) were used reliably (i.e., identically rated or left unchecked) by the assessors across all four replicates. If treating the data as checked/unchecked, the global reproducibility index (RI_i^{FREQ}) took values between 0.77 and 0.49 (mean = 0.66).

In evaluating such values it should be pointed out that, unlike previous studies, they are obtained across >2 replicates, making it less likely for the term $des_{com\,j}$ to equal 1 by chance alone. In order to account for the higher number of replications, one simple solution might be to compute separate reproducibility indices for each couple of replicates, and then divide by the number of combinations. As expected, considering such "averaged" reproducibility indices, significantly higher values – ranging from 0.66 to 0.81 (mean = 0.73) for RI_i^{RATA}, and from 0.72 to 0.88 (mean = 0.79) for RI_i^{FREQ} – were obtained (Table 2).

Inspecting Table 2a and b together we can observe that assessors' reproducibility indices are always lower for intensities than simple frequencies (as it should be), but the interesting result is that the values in Table 2a are in most cases identical or very close to those in Table 2b. This seems to suggest that assessors' reproducibility in a RATA task is comparable to that of a simple checklist, despite the addition of an intensity rating step. Here, let us stress that this conclusion is based on a simple re-analysis of the same

Table 2
Assessors' reproducibility index (Eq. (2)), global as well as by samples, considering frequencies (a) or actual scale points (b).

(a)											
	RI_i^{RATA}										
	Global (4 reps.)	Global (Avg.)[a]	A	B-1	B-2	C	D	E	F	G	H
A1	0.66	0.81	0.68	0.72	0.70	0.68	0.63	0.55	0.66	0.68	0.68
A2	0.48	0.68	0.61	0.54	0.51	0.49	0.43	0.48	0.45	0.40	0.49
A3	0.64	0.79	0.61	0.60	0.68	0.66	0.58	0.63	0.71	0.63	0.64
A4	0.46	0.67	0.66	0.35	0.46	0.43	0.41	0.32	0.49	0.57	0.46
A5	0.63	0.76	0.69	0.61	0.71	0.57	0.60	0.62	0.57	0.68	0.65
A6	0.54	0.71	0.62	0.62	0.54	0.55	0.54	0.45	0.48	0.55	0.52
A7	0.45	0.66	0.48	0.45	0.38	0.46	0.38	0.40	0.74	0.46	0.37
A8	0.56	0.72	0.58	0.49	0.51	0.52	0.49	0.57	0.77	0.54	0.60
A9	0.47	0.67	0.57	0.49	0.32	0.43	0.49	0.46	0.60	0.46	0.40
A10	0.57	0.75	0.58	0.63	0.60	0.54	0.57	0.60	0.55	0.52	0.57
A11	0.63	0.78	0.65	0.62	0.71	0.62	0.66	0.57	0.72	0.55	0.65
Mean	0.55	0.72	0.61	0.56	0.56	0.54	0.52	0.51	0.61	0.55	0.55
(b)											
	RI_i^{FREQ}										
	Global (4 reps.)	Global (Avg.)[a]	A	B-1	B-2	C	D	E	F	G	H
A1	0.68	0.82	0.69	0.72	0.74	0.69	0.66	0.57	0.68	0.68	0.68
A2	0.53	0.75	0.66	0.60	0.57	0.52	0.51	0.52	0.51	0.41	0.51
A3	0.69	0.83	0.66	0.63	0.72	0.72	0.63	0.66	0.77	0.69	0.69
A4	0.54	0.75	0.73	0.43	0.55	0.49	0.44	0.36	0.58	0.68	0.55
A5	0.68	0.82	0.69	0.68	0.75	0.62	0.65	0.69	0.63	0.71	0.71
A6	0.58	0.77	0.68	0.68	0.58	0.63	0.60	0.45	0.54	0.58	0.55
A7	0.50	0.73	0.52	0.49	0.45	0.51	0.42	0.42	0.78	0.55	0.43
A8	0.63	0.80	0.58	0.58	0.60	0.62	0.52	0.62	0.82	0.68	0.69
A9	0.49	0.72	0.58	0.54	0.35	0.43	0.55	0.51	0.60	0.48	0.42
A10	0.59	0.78	0.63	0.63	0.65	0.54	0.58	0.60	0.60	0.55	0.58
A11	0.77	0.88	0.77	0.78	0.83	0.71	0.80	0.69	0.82	0.72	0.82
Mean	0.61	0.79	0.65	0.61	0.62	0.59	0.58	0.55	0.67	0.61	0.60

[a] Denotes the mean of reproducibility indices computed separately for each couple of replicates.

Table 3
Attribute stability index (Eq. (3)) for all evaluated attributes, considering frequencies (SI_k^{FREQ}) or actual scale points (SI_k^{RATA}).

Attribute	SI_k^{FREQ} (%)	SI_k^{RATA} (%)	Attribute	SI_k^{FREQ} (%)	SI_k^{RATA} (%)
Appearance			Old nuts	78.8	78.8
Glossy	55.6	35.3	Cacao	42.4	23.2
Homogeneous	56.6	40.4	Caramel	62.6	62.6
Matt	56.6	46.4	Dark Beer	64.6	63.6
Odour			Nuts	83.8	83.8
Floral	71.7	70.7	Off-flavors	85.8	85.8
Burned	64.6	64.6	Roasted	53.5	52.5
Caramel	77.7	77.7	Beetroot	92.9	92.9
Fermented	81.8	81.8	Smoke	63.6	62.6
Honey	58.6	57.6	Butter	66.7	65.6
Hay	46.5	34.3	Acidic	42.4	40.4
Cacao	37.4	21.2	Sweet	49.4	39.4
Dark bitter	49.5	49.5	Tobacco	57.6	57.6
Dark berries	81.8	81.8	Dried fruit	43.4	40.4
Cardboard	73.7	73.7	Vanilla	73.7	72.7
Roasted	77.7	75.6	Yoghurt	56.6	55.5
Red berries	87.9	87.9	*Texture*		
Acidic	49.5	49.5	Astringent	39.4	27.3
Sweet	51.5	43.4	Soft	40.4	29.3
Tobacco	47.5	46.5	Coating	51.5	49.5
Wood	52.5	52.5	Gritty	72.7	66.7
Dry ash	66.7	66.7	Hard	58.6	30.3
Dried Fruit	55.5	55.5	Melting	50.5	14.1
Vanilla	86.9	86.9	Tart	70.7	70.7
Yoghurt	85.8	85.8	*Aftertaste*		
Flavour/Taste			Bark	76.8	76.8
Balsamico	81.8	78.8	Bitter	53.5	28.3
Cheap chocolate	68.7	64.7	Burned	55.5	52.5
Bitter	54.5	26.3	Cacao	27.3	15.1
Burned	50.5	46.5	Cacaomilk	50.5	48.5
Burned roasted onions	81.8	77.8	Roasted	61.6	60.6
Berry-like	57.6	54.5	Acidic	39.4	37.4
Deep roasted coffee	51.5	48.5	Sweet	47.4	35.3
Fermented	67.7	65.6	Sweet Cream	59.6	51.5
Cream	42.2	42.2	Tobacco	62.6	61.6

dataset, not on an actual comparison of CATA and RATA, so empirical evidence is still needed to support this claim. Furthermore, while it might be reasonable to carry out such comparison within the present experimental conditions (with a panel of semi-trained assessors who were familiar with the ballot and the product set), we expect that the difference between CATA and RATA would be larger with a panel of naïve consumers who are known to adopt slightly different cognitive strategies depending on which of these two formats is used (Ares et al., 2014b).

Finally, assessors' reproducibility did not vary notably from one sample to another (see breakdown in Table 2), suggesting that the sensory complexity of the focal samples was approximately the same, and that differences in performance were more related to individual capabilities.

3.2. Attribute stability

The second aim of the study pertained to reproducibility in the use of individual attributes.

Stability indices obtained for individual attributes (SI_k) are reported in Table 3. The average values were 55.8 for SI_k^{RATA}, and 61 for SI_k^{FREQ}.

Table 3 shows that stability indices for individual attributes span a rather large range of values ($27.3 \leqslant SI_k^{FREQ} \leqslant 92.9$; $14.1 \leqslant SI_k^{RATA} \leqslant 92.9$), indicating that some attributes were more problematic than others. Nevertheless, most attributes (50/65 if considering frequencies, 38/65 if considering intensities) were reproducible at or above 50%. Conversely, very few attributes showed truly low reproducibility, e.g. inspecting Table 3 shows there to be nine attributes with SI_k^{RATA} less than 30%, and only one with SI_k^{FREQ} less than 30%.

At face value, these results indicate a relatively good attribute stability since the odds of using the same attribute identically across 4 replicates (the term sam_i in Eq. (3)) by chance alone are 6.25% (1 out of 16) or 0.4% (1 out of 256) if considering frequencies or intensities respectively. Moreover, if recomputing SI_k considering only two replicates at a time (as done for assessors), significantly higher values were obtained: the mean SI_k value was 72.9 and the range was 34.3–98.9.

Again, attribute stability indices are always lower for RATA than simple frequencies, but Table 3 shows that the two values are identical or very close for the vast majority of attributes, suggesting that, within the conditions tested, the reproducibility of RATA might be comparable to that of a CATA task. A possible exception to the latter statement pertained to appearance and texture attributes, where a larger proportion of major deviations (compared to other sensory modalities) between the two values was observed (Table 3). It is unclear whether this is a pattern or a spurious result, and future studies on multiple product categories might be able to shed light on that.

3.3. Reproducibility of sensory characterizations

The last evaluative criteria pertained to the stability of sample configurations obtained from individual replicates. Fig. 1 shows the sample space obtained in first two dimensions on the MFA

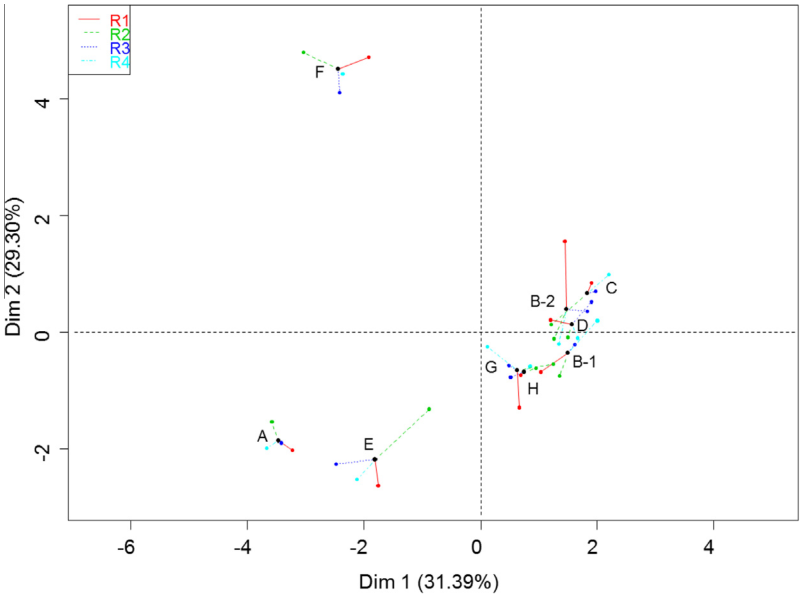

Fig. 1. First and second MFA dimensions showing consensus sample configuration with superimposed partial configurations from individual replicates.

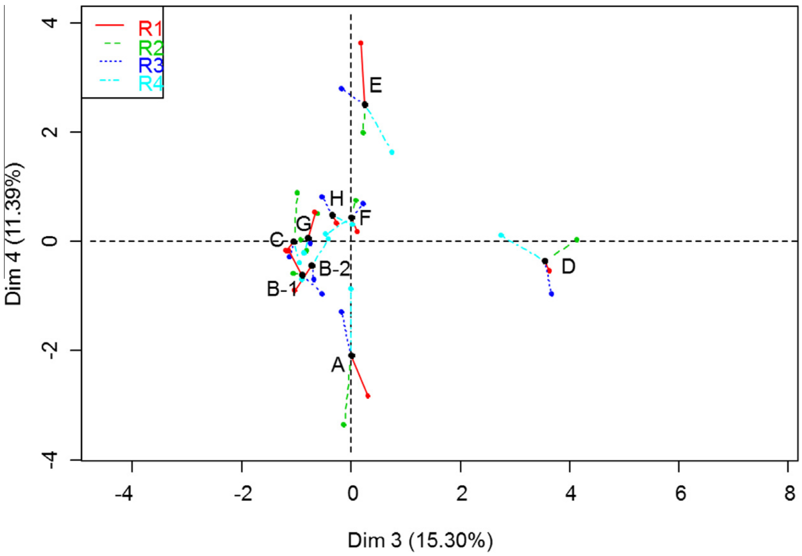

Fig. 2. Third and fourth MFA dimensions showing consensus sample configuration with superimposed partial configurations from individual replicates.

model (four dimensions retained, 87% total variance explained). We do not show here a full analysis of the correlation of original attributes with the MFA dimension, since a detailed sensory profile of the focal samples is outside of the scope of this paper. However, let us briefly report that the first MFA dimension mostly described variation with regards to bitterness and intensity of cacao flavor and aftertaste (negatively correlated sample A, the sample with lowest cocoa content), whereas the second dimension described variation with regards to burned flavor and aftertaste, roasted coffee flavor, and burned roasted onion flavor. The only sample that really correlated with the second dimension was Sample F (the over-roasted sample) suggesting that this sample was very different from the rest and hence that overly roasting the cocoa beans was a very impactful production defect from a sensory perspective.

The third and fourth dimensions (Fig. 2) respectively explained variation with regards to matt appearance (associated to Sample D, the "bloomed" defect), and to textural softness (associated to Sample E, a chocolate with too much lecithin).

As shown in Figs. 1 and 2, although some minor differences in the positions of the samples can be seen from one replicate to another, the conclusions regarding similarities and differences between samples did not change. The two blind duplicates (B-1 and B-2) obtained very similar scores on all MFA dimensions considered (Respectively: 1.5 and 1.5 on Dimension 1; −0.3 and −0.4 on Dimension 2; −0.8 and −0.7 on Dimension 3; −0.6 and −0.5 on Dimension 4) and indeed they are located very close to each other in both Figs. 1 and 2.

Supporting this visual interpretation, the RV coefficients between configurations obtained from individual replicates ranged from 0.81 to 0.97 (Mean = 0.87) when considering two MFA dimensions, and between 0.79 and 0.91 (Mean = 0.83) when considering four MFA dimensions, indicating a high degree of configurational similarities (Table 4). In the sensory descriptive analysis literature, RV values ranging from 0.85 (e.g., Lawless & Glatter, 1990) to 0.95 (e.g., Blancher, Clavier, Egoroff, Duineveld, & Parcon, 2012) are generally considered to indicate good reproducibility across replicates

Table 4
RV coefficients quantifying the similarities between sample configurations from individual replicates considering 2, 4 or all MFA dimensions.

	RV (2 MFA dims.)	RV (4 MFA dims.)	RV (All MFA dims)
R1 vs. R2	0.81	0.81	0.70
R1 vs. R3	0.90	0.86	0.76
R1 vs. R4	0.87	0.81	0.66
R2 vs. R3	0.85	0.80	0.81
R2 vs. R4	0.85	0.79	0.73
R3 vs. R4	0.97	0.91	0.85

from the same panel. Thus, the results presented in Table 4 clearly indicate that RATA is a reproducible method, especially when considering dimensions that are typically of interest to the sensory professional (two or four). RV values were slightly lower when considering the full MFA model with eight dimensions (Mean = 0.75). Since nearly 90% of the variance is explained in the first four dimensions, this probably indicates that higher order dimensions modeled mostly noise.

As a general point, we should note that configurational similarities are strongly dependent on sample size and product differences. Therefore, additional studies are advised to assess whether the same conclusions obtained in this paper are generalizable to different combinations of sample size and product differences.

4. Conclusion

This work has examined the reproducibility of results obtained from rate-all-that-apply (RATA) questions with semi-trained assessors, using sensory assessment of common defects in chocolate production as a case study.

The main findings were that, although within-assessors reproducibility was moderate, at panel level RATA showed a very good reproducibility, as indicated by the high configurational agreement between product maps obtained from individual replicates. The method showed a good reproducibility also at the level of individual attributes with assessors using, on average, more than 55% of the attributes identically across all four replicates.

It should be observed that the reproducibility indices considered in this work (Eqs. (2) and (3)) might be too harsh when considering more than two replicates. As a simple workaround, we proposed to consider individual combinations of replicates individually and to average the values obtained across assessors (for RI_i) or attributes (for SI_k), in order to obtain a more balanced perspective. Future work might consider more efficient solutions, such as developing indices that take *partial* reproducibility into account.

Interestingly, the reproducibility indices considered did not vary substantially when treating the data as simple frequencies versus when considering actual intensity ratings. Therefore, we speculate that the reproducibility of RATA with semi-trained assessors might be comparable to that of a simple checklist, although further evidence in support of this claim is needed.

In conclusion, these results indicate that RATA is a valid and reliable sensory profiling tool. Its inherent characteristics make it particularly advantageous in industrial contexts where small semi-trained panels (e.g. of co-workers) are most readily available, but where the time or the budget for sensory evaluation is often limited.

Acknowledgements

We thank Julie Boeck Ogstrup and the Toms Confectionery Group for supporting this project and providing guidance with regards to the sample preparation. The company employees that took part in the sensory panel are also gratefully acknowledged. We acknowledge further support from the Department of Food Science, University of Copenhagen, Denmark, with which both authors were affiliated at the time this work was initiated.

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 *Proceedings of the 7th pangborn sensory science symposium, Hyatt Regency, Minneapolis, MN, USA*.

Ares, G., Antúnez, L., Giménez, A., Roigard, C. M., Pineau, B., Hunter, D. C., et al. (2014a). Further investigations into the reproducibility of check-all-that-apply (CATA) questions for sensory product characterization elicited by consumers. *Food Quality and Preference, 36*, 111–121.

Ares, G., Bruzzone, F., Vidal, L., Cadena, R. S., Giménez, A., Pineau, B., et al. (2014b). 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., & Jaeger, S. R. (2013). Check-all-that-apply questions: Influence of attribute order on sensory product characterization. *Food Quality and Preference, 28*, 141–153.

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

Ares, G., Tárrega, A., Izquierdo, L., & Jaeger, S. R. (2014c). 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.

Beckett, S. T. (2009). Conching. In S. T. Beckett (Ed.), *Industrial chocolate manufacture and use* (4th ed.. York, UK: Wiley-Blackwell.

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*, 36–43.

Briones, V., & Aguilera, J. M. (2005). Image analyses of changes in surface color of chocolate. *Food Research International, 38*, 87–94.

Campo, E., Do, B. V., Ferreira, V., & Valentin, D. (2008). Aroma properties of young Spanish monovarietal white wines: A study using sorting task, list of terms and frequency of citation. *Australian Journal of Grape and Wine Research, 14*, 104–115.

Civille, G. V., Carr, B. T., & Munõz, A. M. (1992). *Sensory evaluation in quality control*. New York: Van Nostrand Reinhold.

Costell, E. (2002). A comparison of sensory methods in quality control. *Food Quality and Preference, 13*, 341–353.

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

ISO 8586–1. (1993). Sensory analysis – general guidance for the selection, training, and monitoring of assessors. Part 1. International Organization for Standardization.

Jaeger, S. R., & Ares, G. (2015). RATA questions are not likely to bias hedonic scores. *Food Quality and Preference, 44*, 157–161.

Jaeger, S. R., Chheang, S. L., Yin, J., Bava, C. M., Gimenez, A., Vidal, L., et al. (2013a). Check-all-that-apply (CATA) responses elicited by consumers: Within-assessor reproducibility and stability of sensory product characterizations. *Food Quality and Preference, 30*, 56–67.

Jaeger, S. R., Giacalone, D., Roigard, C. M., Pineau, B., Vidal, L., Giménez, A., et al. (2013b). Investigation of bias of hedonic scores when co-eliciting product attribute information using CATA questions. *Food Quality and Preference, 30*, 242–249.

Lawless, H. T., & Glatter, S. (1990). Consistency of multidimensional scaling models derived from odor sorting. *Journal of Sensory Studies, 5*, 217–230.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software, 25*, 1–18.

Meyners, M., & Castura, J. C. (2014). Check-all-that-apply questions. In P. Varela & G. Ares (Eds.), *Novel techniques in sensory characterization and consumer profiling* (pp. 271–305). Boca Raton: CRC Press.

R Development Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

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. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 25*, 257–265.

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

Worch, T., & Piqueras-Fiszman, B. (2015). Contributions to assess the reproducibility and the agreement of respondents in CATA tasks. *Food Quality and Preference, 40*, 137–146.