Acta Alimentaria, Vol. 43 (1), pp. 61–75 (2014)
DOI: 10.1556/AAlim.43.2014.1.7

COMPARISON OF CONSENSUS PROFILES OBTAINED AT THE END OF PRODUCT-SPECIFIC TRAINING WITH PROFILES OBTAINED BY INDIVIDUAL MEASUREMENTS AND STATISTICAL ANALYSIS

M.M. Rodriguez[a, b], M. López Osornio[b]*and G. Hough[b]

[a] UNICEN, Av. Del Valle 5737, B7400JWI, Olavarría, Buenos Aires. Argentina
[b] Experimental Institute of Food Technology, Scientific Research Commission of the Province of Buenos Aires, Hipolito Yrigoyen 931, (B6500) 9 de Julio. Argentina

(Received: 25 June 2012; accepted: 5 November 2012)

In many occasions descriptive analysis consists of product-specific training where the samples to be measured are used during the training. Towards the end of the training period it is common practice to present these samples and reach a consensus on their profiles, which we have called Training Consensus Profiles (TCP). Following the TCP, the samples are scored by each assessor and the results are statistically analysed to obtain statistical profiles. The objective of the present work was to compare the TCP with the statistical profiles in samples from three different food categories: fernet (an herb-based alcoholic drink), mayonnaise, and spaghetti. General Procrustes analysis showed that the TCP and statistical profiles were similar. A case is made, that if this type of training and measurement are to be followed, the statistical measuring stage could be left aside, directly reporting the results obtained from the TCP. Advantages and limitations on reporting these TCP profiles are discussed.

Keywords: sensory evaluation, descriptive tests, consensus, fernet, mayonnaise, spaghetti

Books on sensory analysis (Lawless & Heymann, 2010; Stone & Sidel, 2004; Meilgaard et al., 2007) present different methods used to perform sensory descriptive analysis. The Flavor Profile® is a consensus technique. Using standardized techniques of preparation, presentation, and evaluation, 4–6 assessors are trained to precisely define flavours of the product category during a 2 to 3 weeks programme. The panel is exposed to a wide range of products within the food category. After this exposure, the assessors review and refine the descriptors; reference standards and definitions for each descriptor are also created during the training phase. At the end of the training phase, the panel has a frame of reference for expressing the intensities of the descriptors. This training phase of the Flavor Profile method is similar to the training phases of other descriptive methods. In Quantitative Descriptive Analysis (QDA) the recommended number of judges is between 10 and 12 and during the training period the panel leader has less influence than in other methods in defining descriptors. The main difference between Flavor Profile and QDA is that after developing a standardized vocabulary to describe the sensory differences among the samples, the samples' profiles are defined by consensus using the Flavor Profile, while for the QDA method the profiles are obtained by statistical analysis after each assessor individually measures the samples. Sensory Spectrum is a descriptive method where assessors do not generate a panel-specific vocabulary; the language used to describe particular products is chosen a priori and remains the same for all products within a category over time. Additionally, the scales are standardized and anchored

* To whom correspondence should be addressed
Phone: +54-2317-425507; fax: (02317) 422305; e-mail: mercedes@desa.edu.ar

with multiple reference points, which usually extends the training period in comparison to QDA. Once the training is completed the samples' profiles are defined in a similar mode to QDA. As pointed out by Lawless and Heymann (2010), QDA and Sensory Spectrum techniques have been adapted in many different ways. Academic researchers frequently employ the general guidelines of these methodologies to evaluate products. To generalize, we shall refer to the Statistical Descriptive Method (SDM) as one where, after an initial training period, each individual assessor measures the samples and the profiles are defined by statistical analysis.

When training a panel for SDM, all or some of the samples to be profiled are presented to the assessors. In the initial sessions these samples are used to generate descriptors. In further sessions they are evaluated to test the adequacy of descriptor definitions and references. In the final training sessions the samples to be profiled are presented to test the final ballot and to decide if the panel are in overall agreement on the use of the descriptors and the corresponding scale. Very often, by the end of the training sessions corresponding to a SDM, the panel leader has a relatively clear idea as to the profile of these samples.

It is also inevitable for the assessors to become familiar over the training period with the sensory properties of the samples to be measured. Experienced assessors will expect to find these samples in the sets they are given to measure in their individual booths. Will they consciously or unconsciously try to reproduce what was discussed over the samples during training? For example, in the profiling of chocolate flavoured milk, the sweetest sample may have also been associated to having the highest viscosity. So when the assessor is measuring the samples and he/she detects a sample with high viscosity, it is very probable that he/she will also score the sample high in sweetness.

The objective of the present work was to compare the profiles obtained by discussion and consensus at the end of the training period with the profiles obtained by individual measurements followed by statistical analysis in samples from three different food categories: fernet (an herb-based alcoholic drink), mayonnaise, and spaghetti. The comparison of this type of profiles has not been reported in the literature.

1. Materials and methods

1.1. Samples and presentation

For each food category between 6 and 8 commercial samples were initially bought at a local supermarket far from their "best-before dates". These were evaluated by the authors with the help of 2 trained assessors. Final choice of samples to be profiled was based on their perceptible sensory differences among them. All products were stored at room temperature in darkness till evaluation.

Four commercial fernets (L, M, N, and P) were chosen. Fernet has an alcohol content of 45%, thus samples were diluted in tap water in the proportion of 1 part of fernet + 1 part of water (ASTM, 2004). As from here it should be understood that samples of fernet are diluted samples. For transparency evaluation, 15 ml of the samples were placed in 4.5 cm glass test tubes and observed by placing the tube at eye level and observing the light coming through the sample. For colour intensity, 20 ml of the samples were evaluated in a 5.5 cm diameter plastic Petri dish. For aroma and flavour 30 ml of fernet was served in a 70 ml capacity wide-mouthed glasses covered with a plastic lid; having the plastic lid allowed assessors to receive the full impact of aroma when uncovering the glass. Bread and water were used as palate cleansers between samples.

Three mayonnaise samples (coded A, B, and C) were evaluated. Their composition as stated on their labels was: sunflower oil (samples A and B: 43% and C: 73%), pasteurized egg, water, sugar, salt, lemon juice, vinegar, carotene, citric acid, potassium sorbate, and xantham gum. Sample B had the same ingredients with the addition of mustard. For appearance evaluation a spoonful of mayonnaise was deposited on a 5.5 cm diameter Petri dish. For aroma and flavour, approximately 20 g were served in 70 ml capacity plastic glasses covered with a plastic lid. To avoid appearance changes, samples were served with a maximum of 10 min previous to evaluation.

Four commercial spaghetti samples were evaluated (E, F, G, and H). Their compositions as stated on their labels were: wheat semolina, powdered egg, ferrous sulphate, niacin, and vitamins B1 and B2. Spaghetti was cooked in a proportion of 100 g of spaghetti per litre of boiling tap water. Cooking time was as recommended on the packages (8–9 minutes, depending on the sample) and was measured as from the time the spaghetti was placed into the boiling water. Heating was regulated to maintain a gentle boil. After cooking, the pasta was drained. Approximately 50 g of spaghetti was served in covered 120 ml thermo foam glasses. Assessors evaluated the samples at 70–75 °C. Rinsing between samples was with tap water, which was odourless and of constant quality.

Both during training and measurement, samples were coded with 3-digit numbers; the codes were changed from one training session to the next. Illumination was with artificial day-type fluorescent lamps.

1.2. Assessors and panel leader

The panel consisted of 9 female assessors aged between 24 and 49 years, selected and trained following the guidelines of the ISO (1993) Standard. Although this standard recommends a minimum of 10 assessors, as all assessors had experience in descriptive analysis of a variety of food products, 9 was considered adequate. They were all female due to recruitment procedures among housewives who have the time off to do sensory panel work. The panel leader had received training following the guidelines of the ISO (2006) Standard.

1.3. Product specific training

For the three product categories training was similar. This stage was carried out with assessors sitting at a round table. In a first session descriptors were generated by assessors evaluating similarities and differences between representative samples. Due to time constraints, texture attributes were not evaluated. For the second and successive sessions the panel leader presented references corresponding to each descriptor. The panel compared the references to the samples to be measured and discussed their validity in representing the corresponding descriptor. The references were scored on the 0 to 10 scales in comparison to the samples to be measured. Positioning of references on the scales was obtained by consensus. The descriptor development stage was completed once the panel felt comfortable with the descriptors and their references in relation to the samples to be measured. Tables 1, 2, and 3 present the descriptors and references for each product category.

Once the descriptor development stage had been completed for a product, an additional 2 sessions per product were used to complete the training by presenting samples whose profiles were discussed with the objective of reaching a consensus among the panel. It is common practice to include samples to be measured during training sessions and discuss their scores. Thus, assessors considered these last sessions as part of their product specific

Table 1. List of descriptors and references of appearance, aroma and flavour of fernet

Attributes	Descriptors	References
Appearance	Light-dark intensity	[a]
	Transparency	[a]
Aroma	Total initial intensity	[a]
	Alcohol	Ethanol 96° (10)[b]
	Sugar cane/candy[c]	33 g of "uruzu" herb (*Rynchosia selecta)* in 500 ml of ethanol 50° (8)
	Carqueja[c]	30 g of "carqueja" herb (*Baccharis crispa)* in 100 ml of ethanol 50° 1 ml of this last solution in 70 ml tap water (4)
	Spicy[c]	Fernet Capri (Pernod Ricard, Buenos Aires, Argentina) (3)
	Earthy[c]	6.6 g of zedoaria seeds (*Curcuma zedoaria*) in 100 ml of ethanol 50° (7)
Flavour	Alcohol	Ethanol 96°(10)
	Bitter	0.5 g of caffeine in 1 l of tap water (3)
	Sweet	8 g sucrose in 1 l of tap water (2)

[a]: References for these descriptors were not considered necessary by the panel; [b]: the numbers in brackets indicate the value of the reference on the 0–10 scale; [c]: The same descriptor and reference were used for aroma and flavour

Table 2. List of descriptors and references of appearance, aroma, and flavour of mayonnaise

Attributes	Descriptors	References
Appearance	Quantity of bubbles on surface[a]	
	Yellow colour [a]	
	Gloss (degree in which the surface is glossy/shiny)[a]	
Aroma	Acid[d]	4 ml of alcohol vinegar[b] (8) for aroma, (6) for flavour[c]
	Egg[d]	6 g of egg yolk from an egg boiled 8 min[b] (10)
	Lemon[d]	10 ml of freshly squeezed lemon juice[b] (3) for aroma (10) for flavour
	Garlic/mustard[d]	40 µL of 15% allyl isothiocyanate solution[b] (10)
Flavour	Oil	Natura mayonnaise (AGD, Aceitera General Deheza, Argentina) (6)
	Sweet	1 g of sucrose[b] (5)
	Glutamate	1.5 ml of 0.4% monosodium glutamate solution[b] (7)
	Salty	0.7 g of NaCl[b] (10)
	Heat	1.5 ml of 0.06% Capsicum solution[b] (10)

[a]: References for these descriptors were not considered necessary by the panel; [b]: in 100 g of basic mayonnaise: sunflower oil 84.2%, pasteurized egg yolk 9.1%, water 5.5%, sucrose 0.6%, salt 0.5%, and sodium benzoate 0.1%. (Santa Cruz et al., 2002); [c]: the numbers in brackets indicate the value of the reference on the 0–10 scale; [d]: The same descriptor and reference were used for aroma and flavour

Table 3. List of descriptors and references of appearance, aroma, and flavour of spaghetti

Attributes	Descriptors	References
Appearance	Yellow colour	Pantone®[a] 120 C (3)[b], Pantone® 122 C (8)
	Green colour	Pantone® 394 C (9)
	Quantity of points	Spaghetti Vizzolini (Kraft Foods SA, Argentina) (8)
Aroma	Broth	Liquid resulting from boiling 200 g of lean pork in 1 l of tap water (8)
	Egg	Homemade pasta[c], (10) for aroma, (7) for flavour
	Flour	Cooked dough[d] (5)
	Cereal	Spaghetti Terrabusi (Kraft Foods SA, Argentina) (6) for aroma (3) for flavour
	Turmeric/saffron	0.075 g of turmeric and 0.0125 g of saffron in 100 ml of water hot (8)
Flavour	Egg	Homemade pasta[c], (10) for aroma, (7) for flavour
	Flour	Cooked dough[d] (5)
	Cereal	Spaghetti Terrabusi (Kraft Foods SA, Argentina) (6) for aroma (3) for flavour

[a]: Pantone® Formula Guide (Pantone Inc.); [b]: the numbers in brackets indicate the value of the reference on the 0–10 scale; [c]: 300 g flour, 146 g beaten eggs, 125 g sunflower oil and 22 ml tap water. The dough was left to rest 20 min, cut with a knife and cooked 7 min in boiling water (100 g pasta/1 l water); [d]: 150 g flour and 80 ml tap water, cooked 8 min in 1 l of boiling water

training and were unaware that the consensus results were going to be used for comparative purposes with the SDM.

For the last two training sessions samples were coded with 3-digit numbers. Assessors received all samples corresponding to the product category simultaneously sitting at a round table. For example, they received three mayonnaise samples. For a single consensus session the number of descriptors was limited. For example, for mayonnaise appearance was covered in the first session and aroma and flavour in the second session. Assessors, individually and in silence, measured all samples on all the descriptors covered by the session. For this, they used a structured intensity line scale which went from 0 (low/none) to 10 (high), marked with each digit between 1 and 10. This type of scale is widely used in sensory analysis (Lawless & Heymann, 2010). The scale had the intensity of the corresponding reference marked on it. References were available for assessors who asked for them. Assessors wrote down each sample's number on a position corresponding to the perceived intensity. For example, for mayonnaise, for each scale/descriptor assessors wrote down 3 numbers, each one corresponding to one of the 3 samples that were measured. Once samples had been measured by all assessors on all the descriptors covered by the session, the panel leader initiated the consensus discussion. On a board, the panel leader wrote down the scores for the 3 samples for a descriptor as called out by each assessor. The panel leader led the discussion on the scores searching for a consensus. Samples were re-evaluated by some or all assessors during the consensus discussion of some descriptors. As expected in a consensus discussion, it was up to the leader to stop the discussion at a point she considered that a consensus had been reached. Once consensus had been reached on one descriptor the procedure was repeated for the other descriptors. We shall refer to the profile obtained at the end of the training sessions as the Training Consensus Profile (TCP). The TCP was reached by consensus, not through statistical calculations, such as ANOVA or averages.

Once the assessors had completed the training sessions, they proceeded to measure the samples by the SDM for a product category. These measurements were made 1 or 2 days after they finished the training sessions. Assessors evaluated the samples in a sensory laboratory equipped with individual booths and day-light type fluorescent lighting of the same characteristics as used during training. Samples were served in random order. Measurements were in triplicate in different sessions. Data collection was done using SoPas (Software para Análisis Sensorial, Luis Secreto, Nueve de Julio, Argentina). Assessors were informed that samples presented during these measurement sessions were not necessarily the same ones presented during the previous training sessions.

General Procrustes analysis (Arnold & Williams, 1986) was used to monitor assessors' performance by analysing their residuals and their relative position on the principal coordinate analysis plot. Analysis of variance (ANOVA) was applied to each descriptor considering assessors as random effect and samples as fixed effect. Normality plots (McConway et al., 1999) and Bartlett's homogeneity of variance tests (Snedecor & Cochran, 1989) were performed to test the data as appropriate for ANOVA. Means were compared using Fisher's least significant difference (LSD) at a 5% significance level (O'Mahony, 1986).

Total number of sessions for each product category was as follows:

– Fernet: 7 training and TCP sessions+3STD measuring sessions;
– mayonnaise: 8 training and TCP sessions+3STD measuring sessions,
– spaghetti: 11 training and TCP sessions+3STD measuring sessions.

Note that all sessions for one product category were finished before starting on the next; that is all the fernet sessions concluded before starting on mayonnaise.

1.4. General Procrustes Analysis (GPA)

In order to compare the sample configurations and descriptor loadings of the TCP with the SDM profiles (mean scores over samples), GPA (Arnold & Williams, 1986) was applied to the sample×descriptor matrixes corresponding to each product category. ANOVA for the SDM and GPA were performed using procedures from Genstat 10th Edition (VSN International Ltd, Hemel Hampstead, UK).

2. Results and discussion

General Procrustes analysis showed that all assessors performed adequately in all three product categories when using the SDM method. Normal plots were approximately linear and Bartlett's test did not show significant variance differences between samples.

2.1. Fernet

Table 4 shows the scores obtained from the TCP and SDM. GPA results (Fig. 1) showed that overall configurations obtained from both profiles were similar. Scores were alike for 14 of the 15 descriptors evaluated. As an illustration, Fig. 2 is an example of when both profiles were equivalent in their final scores, in this case for sugar cane/candy flavour. Figure 3 is an example of when both profiles were not equivalent, in this case for bitterness. Samples N and L had similar scores; however, for this descriptor samples M and P were separated by the TCP but not by the SDM. For carqueja aroma, samples M and N received scores of 0 for the TCP; yet the average scores for these samples in the SDM were slightly above 0. Thus, some of the

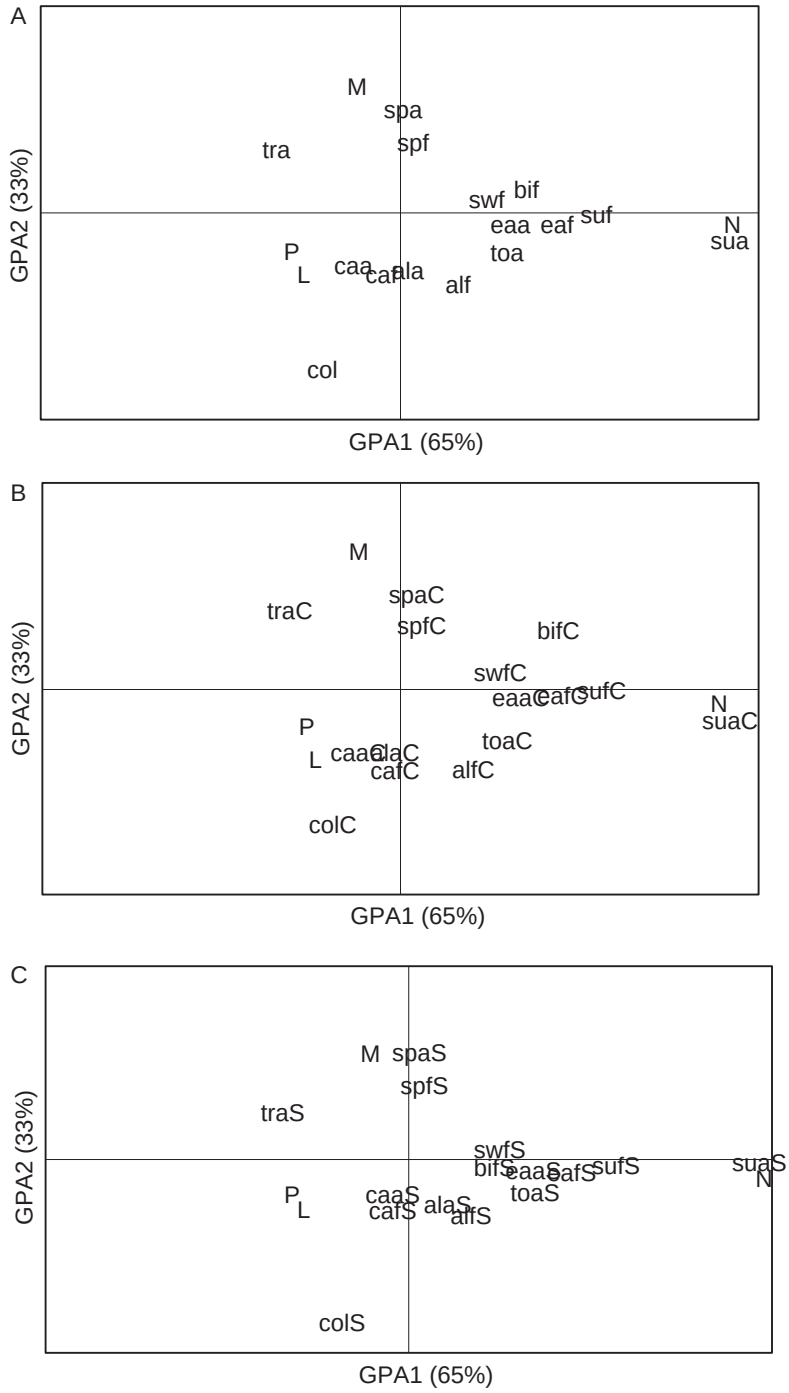

Fig. 1. Fernet general Procrustes analysis (GPA): A: GPA consensus configuration; B: GPA configuration of training consensus profile (T); and C: GPA configuration of statistical descriptive method profile (S). Upper case letters refer to samples and lower case letters to descriptors (Table 4)

assessors in some of the repetitions found a presence of carqueja aroma when they had agreed during training that these samples did not have this aroma. On all descriptors, the range of TCP scores across the samples was higher than that of the SDM average scores.

Table 4. Trained consensus profiles and mean scores of statistical profile for appearance, aroma, and flavour of fernet

Attributes	Descriptors	Trained consensus profile				Statistical profile				$LSD_{0.05}$
		L	M	N	P	L	M	N	P	
Appearance	Colour (col)	10	4.5	6	9	9.2	4.8	6	9.1	0.9
	Transparency (tra)	6	9	4	8	7.3	8.2	4.6	8.3	0.9
Aroma	Total Intensity (toa)	6.5	5	8	7	6.4	5.8	7.9	6.2	0.5
	Alcohol (ala)	6	3	4	5	5.3	4.1	5	5.3	0.9
	Sugar cane (sua)	0.5	0	7	0.5	0.5	1.3	7	0.6	0.9
	Carqueja (caa)	2.5	0	0	2.5	1.7	0.5	0.3	1.9	0.7
	Spicy (spa)	0	3	0	0	0	2.5	0	0.3	0.4
	Earthy (eaa)	0	0	2	0.5	0.5	0.3	2	0.3	0.3
Flavour	Alcohol (alf)	7	4	7	6.5	6.7	5.2	6.8	6.4	1.3
	Bitter (bif)	5.5	7	8	4	5.7	5.8	7.5	6.2	1.1
	Sweet (swf)	0	1	2	1	0.5	0.8	1.6	0.5	0.6
	Sugar cane (suf)	0	0.5	4	0.5	0.4	0.8	4	0.5	0.7
	Carqueja (caf)	3	0	1	2	1.9	0.4	0.4	1.9	0.5
	Spicy (spf)	0	2	0	0	0	1.7	0	0.2	0.3
	Earthy (eaf)	0	0	3	0	0.6	0.5	3	0.5	0.5

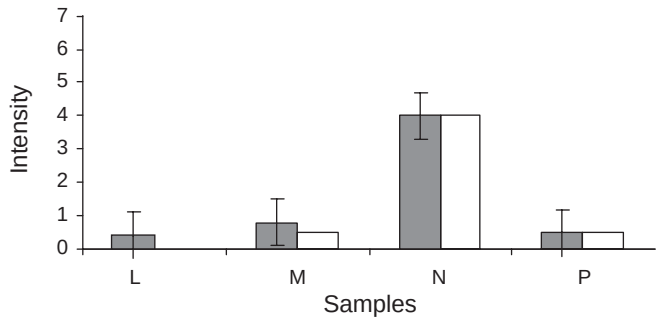

Fig. 2. Training consensus profile (TCP) and statistical descriptive method profile (S) for sugar cane/candy flavour of fernet. Vertical bars indicate the least significant differences (P≤0.05). ▩: Statistical; □: consensus

2.2. Mayonnaise

Table 5 shows the scores obtained from the TCP and SDM. GPA results (Fig. 4) showed that overall configurations obtained from both profiles were similar. For egg flavour, samples A and B had equivalent scores; for sample C the values were 0.5 and 1.8 for the TCP and

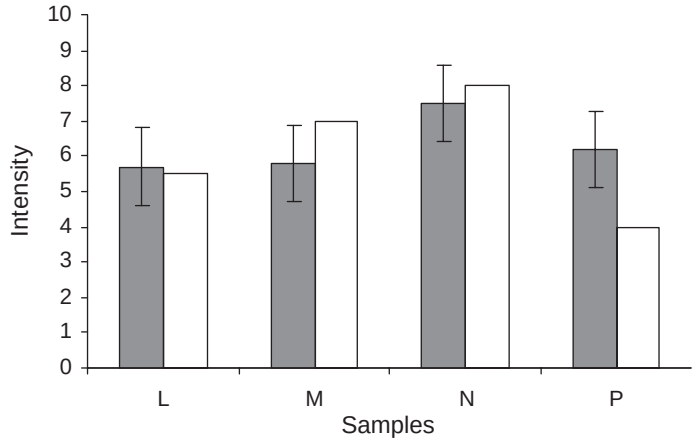

Fig. 3. Training consensus profile (TCP) and statistical descriptive method profile (S) for bitter flavour of fernet. Vertical bars indicate the least significant differences (P≤0.05). ■: Statistical; □: consensus

SDM, respectively. A similar pattern was observed for lemon flavour. Whether these differences are of practical importance would depend on the specific reason the descriptive analysis was called for. Sample C had an average score of 1.8 given by the SDM for egg flavour. This could have been a consequence of some assessors confusing this sample with sample A. Sample A had no acid flavour in the TCP, yet for the SDM the average was 1.7. On the other hand, sample B had scores of 6 and 4.4 for the TCP and SDM, respectively. It would seem that assessors were more conservative when measuring in the SDM: they increased the level of sample A and decreased the level of sample B. The result was a decrease in the score ranges for the SDM in relation to the TCP. This phenomenon was observed for other descriptors: for 14 of the 17 descriptors the range of TCP scores was higher than the range of SDM average scores; for the other 3 descriptors, the difference in ranges was less than or equal to 0.4.

Table 5. Trained consensus profiles and mean scores of statistical profile for appearance, aroma, and flavour of mayonnaise

Attributes	Descriptors	Trained consensus profile			Statistical method			$LSD_{0.05}$
		A	B	C	A	B	C	
Appearance	Bubbles (bub)	7	3	7	6.9	4.3	6.9	0.5
	Yellow color (col)	3	5.5	7	3.7	6	7.9	0.5
	Gloss (glo)	7.5	7	8	7.5	7.3	7.4	NS
Aroma	Acid (aca)	0	4	6	1.5	3.8	4.5	0.7
	Egg (ega)	5	0	1	4.5	0.4	1.2	0.4
	Lemon (lea)	4	0	0	3.5	0.3	0.6	0.4
	Garlic/mustard (gaa)	0.2	5	0	0.2	4	0.2	0.6

Table 5. Continued

Attributes	Descriptors	Trained consensus profile			Statistical method			$LSD_{0.05}$
		A	B	C	A	B	C	
Flavour	Acid (acf)	0	6	3	1.7	4.4	3.2	0.7
	Egg (egf)	4	0	0.5	4.1	0.3	1.8	0.4
	Lemon (lef)	5	0	0.5	3.9	0.2	1.6	0.5
	Garlic/mustard (gaf)	0	6	0	0.2	4.9	0.1	0.4
	Oil (oif)	3	0	6	2.8	1.1	5	0.7
	Sweet (swf)	2	0	0	1.8	0.2	0.6	0.3
	Glutamate (glf)	0	6	0	0.4	4.4	0.2	0.3
	Salty (saf)	1.5	3	5	2.2	3.4	4	0.5
	Heat (hef)	0	3	0	0.1	3.5	0.4	0.4

NS: Non-significant

2.3. Spaghetti

Table 6 shows the scores obtained from the TCP and SDM. GPA results (Fig. 5) showed that overall configurations obtained by both methods were similar. Cereal aroma values differed for sample F. In the TCP sample F had the highest cereal aroma score, yet this was not recognized in the SDM. To a lesser degree this phenomena also occurred for cereal flavour. As observed for some of the fernet and mayonnaise descriptors, the SDM profiles tended to flatten out in relation to the TCP profiles. For 10 of the 11 spaghetti descriptors the range of TCP scores was higher than the range of SDM average scores; for the remaining descriptor (turmeric/saffron) the ranges were equal.

Table 6. Trained consensus profiles and mean scores of statistical profiles for appearance, aroma, and flavour of spaghetti

Attributes	Descriptors	Trained consensus profile				Statistical method				$LSD_{0.05}$
		E	F	G	H	E	F	G	H	
Appearance	Yellow colour (yel)	2	7	1.8	5	4.5	6.1	3.4	3.8	1.4
	Green colour (gre)	4.5	0	0	0	3.7	0.1	0.4	0.3	1.2
	Points (poi)	0.5	0.5	5	8	0.7	1.2	4.8	7.1	1.0
Aroma	Broth (bra)	0	0	2	0	0	0.1	0.5	0.2	0.3
	Egg (ega)	0	0	3	0	0.2	0.8	1.2	0.8	0.5
	Flour (fla)	0	0	2	2	0.3	1.2	1.9	1.4	0.7
	Cereal (cea)	0	6	0.5	4	0	3.3	2	3.5	0.9
	Turmeric/saffron (tua)	6	0	0	0	6	0	0.3	0	0.5
Flavour	Egg (egf)	0	0	2	0	0.1	0.8	1	0.7	0.5
	Flour (flf)	3	1	4	1	2.4	1.7	2.1	1.7	0.6
	Cereal (cef)	0	3	0	2	0	1.6	1.2	1.9	0.6

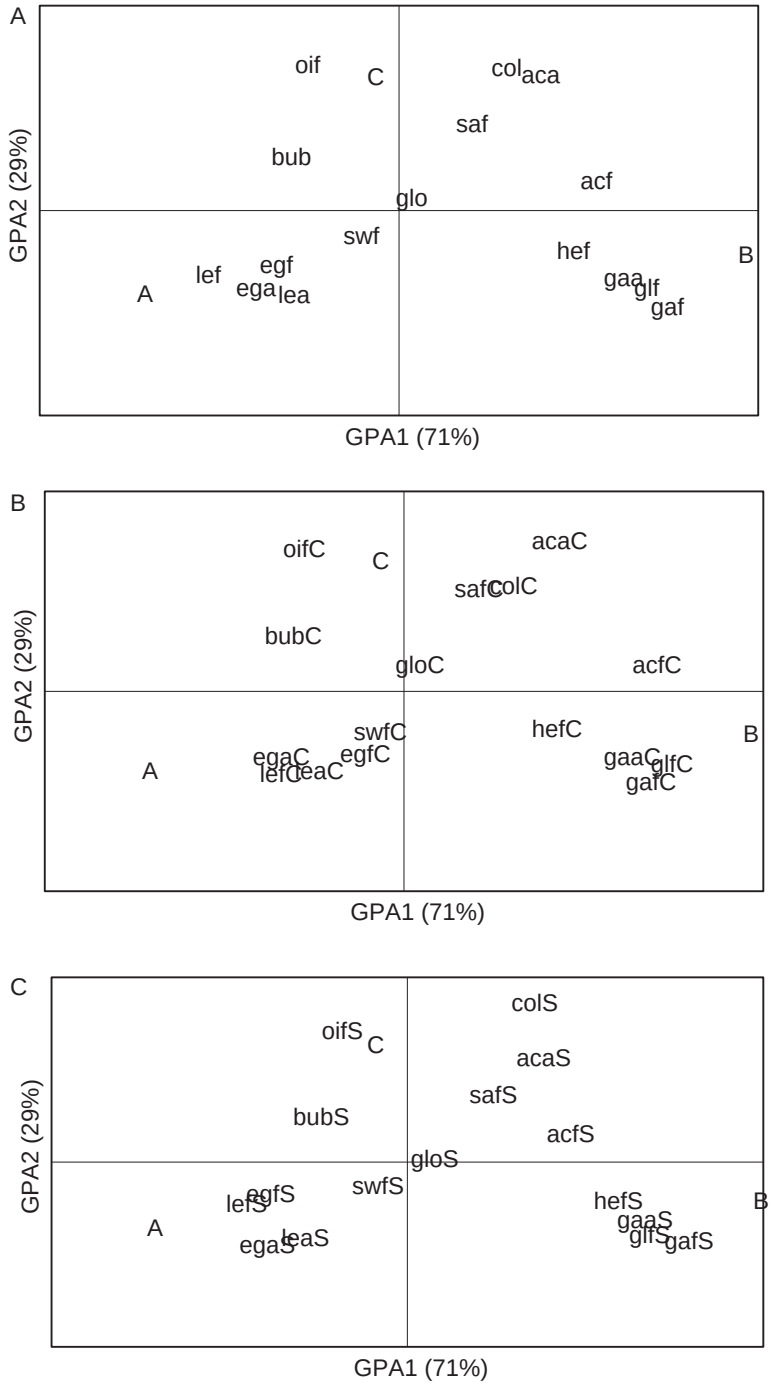

Fig. 4. Mayonnaise general Procrustes analysis (GPA). A: GPA consensus configuration, B: GPA configuration of training consensus profile (T), and C: GPA configuration of statistical descriptive method profile (S). Upper case letters refer to samples and lower case letters to descriptors (Table 5)

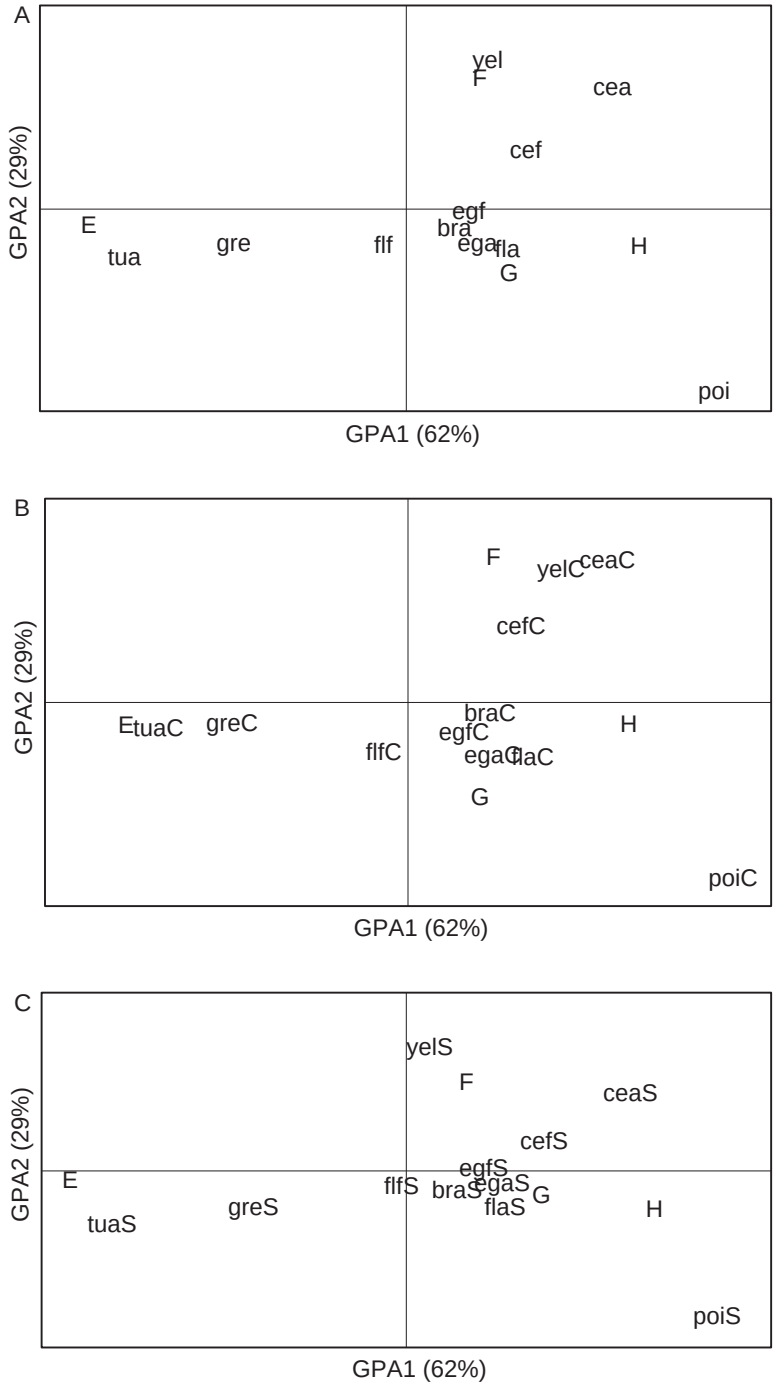

Fig. 5. Spaghetti general Procrustes analysis (GPA). A: GPA consensus configuration, B: GPA configuration of training consensus profile (T), and C: GPA configuration of statistical descriptive method profile (S). Upper case letters refer to samples and lower case letters to descriptors (Table 6)

The overall configurations given by the TCP and SDM profiles were similar for all three product categories (Figs 1, 4, and 5). This showed that assessors, whether discussing scores at the end of product-specific training sessions or measuring samples individually by the SDM, drew up equivalent profiles.

Looking at descriptors individually showed some differences in the profiles. The range of scores obtained for a sample set was generally higher for the TCP, resulting in more pronounced sample differences. The reasons for this can be attributed to the following:

2.3.1.Simultaneous versus monadic presentation of samples. In the TCP each descriptor was discussed with all samples present simultaneously. This could have helped assessors highlight the differences, obtaining a wider spread of the scores over the scale. The ISO (2003) Standard indicates that simultaneous presentation facilitates the comparison of samples. Chambers and Wolf (1996) refer to a "timidity" error, which is the tendency some assessors have of using a limited range of a scale. This error may be enhanced by the SDM, where assessors were presented with one sample at a time, and not being able to contrast it with other samples, tended to measure all samples in a reduced range. Mazzucchelli and Guinard (1999) found that assessors discriminated better and were more reproducible when evaluating samples simultaneously in comparison to monadic presentation. Park and co-workers (2007) found that in a rank-rating procedure, similar to simultaneous sample presentation, there were fewer errors than with a monadic presentation of samples. In this work we chose a monadic presentation in the SDM as it is the customary way of performing this type of measurements; however, further research would call for a comparison of TCP with SDM using simultaneous presentation of samples.

2.3.2. Lack of confidence. In the TCP, assessors have the opportunity to contrast their responses to those of the other members of the panel. For example, if an assessor tastes a sample of mayonnaise and finds that its oil-flavour score is 0, he/she may have doubts about this; among other things because he/she knows that an important mayonnaise ingredient is oil. If in the consensus discussion, other members of the panel also express that they find the sample lacking in oil flavour then this reinforces the assessors' confidence. When the assessor is alone in the booth, a number of questions may arise when confronted with a sample with low oil flavour: "is this the sample that during training had an intensity of 0?", "am I confusing it with that other sample that had an intensity of 3?", "is it a sample not presented during training?" The outcome of asking these questions could readily be to score the sample with a 1 instead of a 0. A similar scenario would occur with a sample scoring high on the scale. Lack of confidence would lead an assessor to score the sample lower in the SDM than in the TCP. The lower range of scores found for the SDM method would be a consequence of the lack of confidence effect. This same effect was recently brought up by Cappuccio in the sensory discussion list (personal communication)*;* his impressions being that in the SDM assessors are somehow afraid to be a "voice out of chorus" and therefore stick to the middle of the scale.

2.3.3. Group dynamics effect:. Syarief and co-workers (1985) cited a 1955 paper by Foster and co-workers (1955) which reported that the round table discussion effect can create qualitative flavour differences where none exist. Syarief and co-workers (1985) followed this up by stating that this problem can be overcome by well-trained panels. Let us add that the panel leader plays a crucial role in the group dynamics. For example, in a panel

of 9 assessors, there may be 6 who agree that a sample of spaghetti has a score of 8 for dark points. The remaining 3 assessors may not agree with this score, but go along with the trend, preferring not to get involved in a discussion with the majority. Thus, the consensus obtained during training would show a score of 8. When the assessors measure the samples in the SDM, they may decide to express what they perceive in the sample and thus, the average score results in a difference in relation to what was obtained by the TCP. However, if this were the case, it would be expected that the resulting SDM average for the sample could either be lower than 8 or higher than 8; and thus the range of scores obtained by the SDM could be lower or higher than the range of scores obtained by the TCP. However, the results of the present work showed that the ranges of the SDM scores were consistently lower. The timidity or lack of confidence errors mentioned above would seem more likely.

In many occasions descriptive analysis consists of product-specific training where the same samples to be measured are used during the training. At the end of the training period it is common practice to present these samples and reach a consensus on their profiles as described above, and we have referred to these as Training Consensus Profiles (TCP). Following the TCP, the samples are scored by each assessor and the results are statistically analysed to obtain the profiles from the SDM. From above it was concluded that under these conditions the TCP and SDM profiles were similar. Thus, a case could be made that if this type of training and measurement are to be followed, the SDM measuring stage could be left aside, directly reporting the results obtained from the TCP. The advantages would be a reduced number of sessions (for our products we had between 7 and 11 training sessions, and 3 STD measuring sessions) and that hardware and software for computerized data entry from descriptive tests would not be required; this last issue would depend on the resources of the laboratory. On the other hand, satisfactory results from the TCP depend on the skills of the panel leader. ISO (2006) Standard provides guidelines on the abilities a panel leader should have. Skills in leadership, group dynamics and communication are highlighted. Training in managing group dynamics, important to obtaining a reliable TCP, is repeatedly mentioned by this Standard. Another caveat to the TCP is that if the number of samples to be profiled is large (for example, 10 samples of beer) it is difficult for assessors to reach consensus on their scores by evaluating them all simultaneously.

Many researchers feel comfortable with SDM due to being able to attach statistical significance to their results. This can be of importance in many cases, such as sensory-instrumental relationships or claim substantiation studies. We are not claiming that SDM should not be used, simply that in some cases the practical conclusions of a study are the same with or without the statistical significance. For example, for the alcohol flavour of fernet (Table 4) the TCP showed that sample M had the lowest level and that the other three samples had similarly higher values with sample P being slightly lower. The SDM profile was similar even though the LSD value attached no significance to the difference between samples M and P. Statistical significance could have been reached with another replication or a higher number of assessors. It should also be noted that when consensus is reached there is a certain degree of implied significance in the reported results. Some assessors might feel that there is a slight difference between two samples, but when the leader asks them if they are sure to be able to find it in another session, or when they see that other assessors are finding a slight difference but in reverse direction, they agree that the samples are very similar in the level of the descriptor and they might as well be considered equivalent. Also, when the panel leader sees that, for example, 6 out of 10 assessors find a consistent difference between two samples, he/she can decide this difference is worth reporting even if not all assessors perceive it.

3. Conclusion

General Procrustes analysis showed that the Training Consensus Profiles and statistical profiles were similar in the evaluation of attributes appearance, aroma and flavour of fernet, mayonnaise, and spaghetti. A case is made, that if this type of training and measurement are to be followed, the statistical measuring stage could be left aside, directly reporting the results obtained from the TCP. However, there are many instances in which the SDM is a necessary procedure.

References

ASTM (2004): *Standard guide for sensory evaluation of beverages containing alcohol.* American Society for Testing Materials, West Conshohocken, E1879 Standard.

Arnold, G.M. & Williams, A.A. (1986): The use of generalized Procrustes techniques in sensory analysis, -in: Piggott, J.R. *Statistical procedures in food research.* Elsevier Applied Science Publishers, London, pp. 233–254.

Chambers, E. & Wolf, M. (1996): *Sensory testing methods*. American Society for Testing Materials, West Conshohocken, pp. 22–23.

Foster, D., Pratt, C. & Schwartz, N. (1955): Variation in flavour judgments in a group situation. *Fd Res., 20,* 539.

ISO (1993): *Sensory analysis – General guidance for the selection, training and monitoring of assessors. – Part 1: Selected assessors.* International Standard Organization, Geneva, No. 8586-1.

ISO (2003): *Sensory analysis – Methodology – General guidance for establishing a sensory profile*. International Standard Organization, Geneva, No. 13299.

ISO (2006): *Sensory analysis, General guidance for the staff of a sensory evaluation laboratory. Part 2: Recruitment and training of panel leaders*. International Standard Organization, Geneva, No. 13300-2.

Lawless, H. & Heymann, H. (2010): *Sensory evaluation of food, principles and practices*. Springer, New York, pp. 153–156, 341–372.

Meilgaard, M.C., Civille, G.V. & Carr, B.T. (2007): *Sensory evaluation technique.* CRC Press, Boca Raton, pp. 173–186.

Mazzucchelli, R. & Guinard, J.X. (1999): Comparison of monadic and simultaneous sample presentation modes in a descriptive analysis of milk chocolate. *J. Sensory Stud., 14,* 235–248.

McConway, K.J., Jones, M.C. & Taylor, P.C. (1999): *Statistical modelling using Genstat.* Arnold Publishers, London. pp. 123–126.

O'Mahony, M. (1986): *Sensory evaluation of food, statistical methods and procedures*. Marcel Dekker, New York, 487 pages.

Park, J., O'Mahony, M. & Kim, K. (2007): Different stimulus scaling errors; effect of scale length. *Fd Qual. Preference, 18,* 362–368.

Santa Cruz, M.J., Martinez, C. & Hough, G. (2002): Descriptive analysis, consumer clusters and preference mapping of commercial mayonnaise in Argentina. *J. Sensory Stud., 17,* 309–325.

Snedecor, G.W. & Cochran, W.G. (1989): *Statistical Methods*. Iowa State University Press, Ames, Iowa, USA. pp. 251–252.

Stone, H. & Sidel, J.L. (2004): *Sensory evaluation practices.* 3rd ed., Elsevier Academic Press, San Diego, pp. 201–244.

Syarief, H., Hamann, D., Giebrecht, F., Young, C. & Monroe, R. (1985): Comparison of mean scores and consensus scores from flavor and texture profile analyses of selected food products. *J. Fd Sci., 50,* 647–650.