Available online at www.sciencedirect.com
ScienceDirect
Food Quality and Preference 19 (2008) 277–285
Food Quality and Preference
www.elsevier.com/locate/foodqual

A comparison of serial monadic and attribute-by-attribute descriptive analysis protocols for trained judges

Rie Ishii [a], Chantal Stampanoni [b,1], Michael O'Mahony [a,*]

[a] Department of Food Science and Technology, University of California, Davis, CA 95616, USA
[b] Givaudan Flavors Ltd., 8600 Dübendorf, Switzerland

Received 3 February 2007; received in revised form 4 August 2007; accepted 7 August 2007
Available online 15 August 2007

Abstract

Trained judges were required to perform a descriptive analysis of a lemon flavored beverage, using a serial monadic protocol and an attribute-by-attribute protocol. They had received sufficient training to establish three intensity exemplars for each attribute in the beverage. They were tested until they reached criterion performance for each protocol. Criterion performance required judges to rate all attributes according to the rank order of their physical strengths. Judges attained 'criterion' performance more rapidly using the serial monadic protocol. The results contrast with those of another study where untrained judges reached criterion performance more rapidly with the attribute-by-attribute protocol.

Keywords: Attribute-by-attribute; Serial monadic; Descriptive analysis; Trained panelists

1. Introduction

The various procedures used for descriptive analysis (Gacula, 1997; Hootman, 1992; Lawless, 1999) require judges to give numerical ratings for chosen sets of sensory attributes of foods or other products. An essential part of this procedure is scaling, the production of the numerical ratings for the attributes. The scaling procedure chosen depends on the particular descriptive analysis procedure being used. It can range from ordinal scaling for the Flavor Profile (Cairncross & Sjöström, 1950; Caul, 1957; Neilson, Ferguson, & Kendall, 1988; Sjöström, Cairncross, & Caul, 1957) through category scales for the Spectrum method (Meilgaard, Civille, & Carr, 1991) and rank-rating or line scales with standards for the Quantitative Flavor Profile (Stampanoni, 1993, 1994) to an unstructured line scale for Quantitative Descriptive Analysis or QDA (Stone & Sidel, 1993, 1998; Stone, Sidel, Oliver, Woolsey, & Singleton, 1974).

The psychophysical literature provides two rival cognitive models for the cognitive processes taking place when using rating scales. One model is supported by Zwislocki and co-workers (Zwislocki, 1983; Zwislocki & Goodman, 1980) and argues that scaling is an 'absolute' process. The rival model supported by Mellers (1983a, 1983b) argues for a relative process. Care must be taken here because the definitions of 'absolute' and 'relative' can vary somewhat; Zwislocki (1983) accused Mellers of not understanding his definition of 'absolute', which appeared to be based on an idea of the number of transformations possible with the data. Here, the absolute model will be understood as regarding the sensation strength elicited by a stimulus attribute as being compared to a set of exemplar sensation strengths stored in memory, each associated with a given numerical value. Stimuli are thus compared to 'absolute' exemplars in memory to obtain numerical ratings, rather than compared with each other. The relative model assumes that the sensation strengths of the stimuli in an experiment are compared with each other and numbers

* Corresponding author. Tel.: +1 530 756 5493.
E-mail address: maomahony@ucdavis.edu (M. O'Mahony).
[1] Present address: Alte Oberdorfstr. 39, 8600 Dübendorf, Switzerland.

0950-3293/$ - see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.foodqual.2007.08.002

assigned accordingly. Essentially, the process is one of ranking the stimuli in order of intensity, while using the numerical estimates to describe the spacing between the ranks.

A calibrated instrument could be said to use a 'cognitive' process that is absolute. Calibration is by definition, the storage of exemplars of intensity strength in memory, each associated with a given numerical value. A calibrated instrument will then compare test samples with its exemplars and will give an appropriate numerical readout. For sensory evaluation, calibration of human judges is rarely used because it is time consuming (O'Mahony & Wong, 1989). Regarding the relative cognitive strategy, a judge who ranks a set of stimuli in order of increasing intensity and then spaces them so that similar intensities are closer together and dissimilar intensities are further apart, may be presumed to be using a relative cognitive strategy. Although the arguments revolving around the absolute vs. relative nature of scaling generally regard the two opposing views as a dichotomy, the two could equally well be regarded as two ends of a continuum, with the possibility of gradual change from one to the other.

As far as descriptive analysis is concerned, it would seem sensible to consider the cognitive strategy that judges use for scaling when deciding the procedure or protocol to be followed. If a judge were to be using a relative cognitive process, then an attribute-by-attribute protocol would be more appropriate. A judge would score each food for a given attribute, while being allowed to re-taste and re-evaluate scores until satisfied that the scores represented the correct 'spacing' between the ranked intensities. The judges would then move on to consider the next attribute and so on until all attributes had been considered. The Flash Profiling method (Delarue & Sieffermann, 2004; Sieffermann, 2000), which could be considered to be a modification of free choice profiling, uses such a protocol. Although expected to be more accurate in terms of scaling errors, such protocol would for many applications be time consuming. There might also be limitations with product preparation (e.g. if 12 samples of ice cream had to be rated on a set of attributes), or the number of samples that can be evaluated simultaneously (e.g. coffee) without a decrease in sensory acuity.

Descriptive analysis is usually performed using a serial monadic protocol. The first food is assessed for the intensities of all its attributes, which are given appropriate scores; then, a second food is assessed in the same way, and a third food, etc. For a judge using a relative cognitive process, an attribute sensation from the first food might be forgotten by the time that attribute was to be assessed for the third or fourth food. The serial monadic protocol would be prone to scaling errors. Therefore, the hidden assumption behind the serial monadic protocol is that the judges should be using an absolute cognitive process for their scaling. If scaling were sufficiently absolute, the attribute intensities of each food would be compared with sets of exemplar intensities in the memory, rather than with each other. If the intensity of an attribute for the first food were to be forgotten while a later food was being assessed, it would not matter.

Ishii, Chang, and O'Mahony (2007) reviewed the various studies that provided evidence for relative and absolute scaling strategies as well as their use in descriptive analysis (eg. Mazzucchelli & Guinard, 1999). They compared the performance of judges using serial monadic and attribute-by-attribute protocols specifically in terms of their skill at scaling. They argued that if descriptive analysis were to be performed using the faster serial monadic protocol, part of the training for the descriptive panel would need to involve the establishment of intensity exemplars in memory. Accordingly, if judges had not established the necessary exemplars, they would achieve 'criterion' performance more rapidly using an attribute-by-attribute protocol. By criterion performance was meant the absence of 'scaling' errors. In other words, the ratings given to each attribute (flavor) would be in the same rank order as the concentrations of the flavorings. The strongest flavoring would be given the highest score, the second strongest flavor, the second highest score and so on. They argued that the extra time required to establish intensity exemplars for the serial monadic protocol would give the advantage of speed to the attribute-by-attribute protocol. They confirmed this and concluded that for untrained judges (consumers), attribute-by-attribute protocols were more suitable.

As the next logical step, it is important to consider what would happen if the experiment of Ishii et al. (2007) were to be repeated with a sample of judges who had already established a set of intensity exemplars. Then, the two protocols might now be equivalent in their simplicity. However, because the time required to perform the monadic protocol would be expected to be less, the total time required to attain criterion performance for the monadic protocol would also be less. Should the attributes be blended into a single product like a beverage, the time required to achieve criterion performance would be mostly taken up with learning to recognize the 'target' attribute strengths in the presence of other attributes, themselves varying in strength.

Accordingly, the goal of this study was to test the hypothesis that judges, who had been trained sufficiently to establish a set of intensity exemplars for each attribute under consideration, would achieve criterion performance more rapidly, using a serial monadic protocol than an attribute-by-attribute protocol.

Judges were first trained to identify the presence or absence of four possible flavors in a lemon beverage: 'citral', 'green', 'ionone beta' and 'eugenol'. They were trained to recognize whether each flavor was in a high concentration, a low concentration or was absent. When there were no more errors, judges were considered to have established three intensity exemplars for each flavor: 'high', 'medium', 'absent'. They were then presented with a set of lemon beverages for which one flavor was in its high concentration, a

second flavor in its low concentration and the remaining two flavors absent (see Table 3). For each lemon beverage the goal of the judges was to reach criterion performance. Judges alternated in their use of serial monadic and attribute-by-attribute protocols and the times taken to achieve criterion performance were noted.

2. Judges' training

2.1. Materials and methods

2.1.1. Stimuli

Stimuli consisted of a base lemon beverage with added flavorings. The base lemon beverage was prepared from a 'lemon syrup' (Givaudan-Roure AG, Dübendorf, Switzerland) which was diluted with purified water (201.1 g syrup, diluted to make 1000 g basic beverage) which itself was made by passing deionized water through a Milli-Q system (Millipore Corp., Bedford, MA). This gave further treatment with ion exchange and activated charcoal, yielding water with conductivity $<10^{-6}$ mho/cm and surface tension >71 dyne/cm.

Additional beverage stimuli were made by adding the following flavors: citral [3,7-dimethyl-2,6-octadienal], green [hex-2,trans-enal,trans-2-hexanal], ionone beta [4(2,6,6-trimethyl-1-cyclonexen-1-yl)3] and eugenol [2-methoxy-4-(2-propenyl)phenol], supplied as concentrates (Givaudan-Roure AG, Dübendorf, Switzerland). The concentrates were diluted by adding 0.1 g of concentrate to 100 g ethanol (95% ethanol 5% water also provided by Givaudan-Roure AG) yielding a flavoring solution. Given weights of flavoring were added to the base beverage to make 1000 g portions of 'flavored beverages'. High and low amounts of flavoring were added to give two strengths for each flavoring, yielding eight flavored beverages (see Table 1). The strengths of the concentrates were determined after preliminary testing by Givaudan-Roure AG.

The nine beverage samples (base lemon beverage and eight flavored beverages) were presented at constant room temperature (18–25 °C) in approx. 10 ml aliquots in 10 ml Pyrex beakers. To ensure that they were odor free, the beakers were washed by rinsing with cold tap water then with hot tap water, were steamed and dried so as to prevent any odors from the beaker interfering with the beverage flavors. Judges tasted the beverages ad lib, deciding for themselves how much to smell or taste, and whether to swallow or expectorate the stimuli. They were allowed to develop the tasting method that worked best for them. This was true for all three stages of screening.

Table 1
Amounts by weight of flavoring added to the base lemon beverage to make 1000 g of 'flavored' beverages for training

Flavoring	High concentration (g)	Low concentration (g)
Citral	7.5	3.0
Green	4.5	1.0
Ionone beta	2.5	1.5
Eugenol	6.0	3.0

2.1.2. Judges

Judges were selected according to a three stage training process. First, judges were required to perform without error, a set of difference tests to discriminate between the base lemon beverage and the beverage with a low level of each added flavor. Then, they had to discriminate between the beverage with the high and low levels of each flavor. For the second stage of training, they had to be able to identify the base lemon beverage and the base lemon beverage with a single flavor added at either a high or low level (total 9 stimuli.) These were presented simultaneously. For the third stage of training, judges had to identify which flavors were present when all the nine beverages were presented serially. After successfully completing these tests, the judges were considered as being able to identify each flavor and estimate whether it was at a high intensity, a low intensity or whether it was absent. Thus, training had allowed them to create three intensity exemplars for each flavoring: 'high', 'low' and 'absent'.

Twelve judges (5M, 7F, age range 20–33 yrs) successfully completed all three stages of training and went on to perform the descriptive analysis. A further 22, not reported here, did not succeed, because of difficulties in detecting the flavorings. All selected judges were students and friends at the University of California, Davis. All but one were naive to sensory testing. All had fasted (except for water) for at least 2 h before testing and were non-smokers.

2.1.3. Procedures

2.1.3.1. Stage 1: difference tests. Judges were first required to discriminate perfectly without feedback between the base lemon beverage and the base with added low concentrations of either citral, green, ionone beta or eugenol, using 6 paired comparisons for each flavor (binomial $p = 0.016$). Having achieved this criterion performance for all four flavors, they then had to show perfect discrimination between beverages with flavors at low and high concentration for each of the four, again with six paired comparisons. Judges who were almost perfect (one or two errors) in their discriminations were given training by being given paired comparisons with feedback to determine whether they could achieve criterion performance without feedback. Once the judge had achieved this for one flavoring, discrimination tests were ceased but continued for other flavorings until criterion performance was achieved for all of them.

They were then given a 'warm-up' procedure (Dacremont, Sauvageot, & Duyen, 2000; Mata-Garcia, Angulo, & O'Mahony, 2007; O'Mahony, Thieme, & Goldstein, 1988; Pfaffmann, 1954; Thieme & O'Mahony, 1990) tasting alternately each of the appropriate stimuli, while knowing their identity. On presentation of the stimuli, if they were easily discriminable, a single sampling of each stimulus was

sufficient; if the stimuli were confusable, further sampling of each was required. For the later discrimination tests (low vs. high concentrations of flavoring), some judges found 'warm-up' not to be necessary. After 'warm-up', judges performed the six paired comparisons, tasting and re-tasting each stimulus as often as desired. Before each set of tests, judges first cleaned the mouth by rinsing at least three times with purified water, and interstimulus rinses were taken ad-lib as desired. The order of presentation was randomized with each of the two stimuli being tasted first for half the tests. The instructions for the discrimination tests were based on the judges' descriptions given during 'warm-up.' Responses were given verbally. This protocol has been called the warmed-up paired comparison (Mata-Garcia et al., 2007; Thieme & O'Mahony, 1990).

2.1.3.2. Stage 2: simultaneous identification. For the second part of the training: simultaneous identification, judges were presented with the base lemon beverage and the eight flavored beverages on a tray. They were required to identify verbally which flavoring was present, and whether it was at a high or low concentration. Judges rinsed at least three times with purified water before the test and could re-taste and re-evaluate the beverages as much as desired. As an aid to memory, they were allowed to make written notes. They rinsed ad-lib before tasting each stimulus. Judges were required to achieve a criterion performance of perfect identification in this task without feedback on two successive occasions. Those who failed were given practice sessions, repeating the task with feedback, until they could achieve criterion performance.

2.1.3.3. Stage 3: serial identification. For the third stage of training: serial identification, judges were presented with all nine beverages in succession and were required to indicate verbally which flavoring, if any, was present. They could re-taste a given beverage as often as desired up to the time they gave their response. Rinsing procedures were the same as in Stage 2. After proceeding to the next beverage, they could not re-taste the former beverage; however, they were allowed to alter their responses. Because of possible context effects and the inability to re-taste a prior beverage, the judges were not required to state whether the flavor was in a high or low concentration. Rinses were taken ad-lib between stimuli.

As before, judges were required to achieve a criterion performance of perfect identification in this task on two successive occasions before training was complete. Again, practice was given where necessary, by having judges perform the task with feedback until they could achieve criterion performance. Once this third task was achieved, training was complete.

2.2. Results

Total training times ranged approx. 1.5–7 h. Judges came for 6–14 sessions. The experimental times required for each of the twelve judges to achieve criterion performance for the three stages of training are given in Table 2. These experimental times involve only actual testing times; other times like establishing rapport are omitted. Judges varied in their session lengths and the amount of a given task they performed; thus, the number of sessions required for each stage of screening was also recorded in Table 2.

There was considerable variation in the amount of time needed to train naive judges to be able to recognize the presence of the four flavorings. The quickest judge (A) required approx. 1.5 h (95 min), while the slowest (L) required approx. 7 h (419 min). Generally, longer training times were caused by a judge having difficulty identifying or detecting one particular attribute. The total testing time (not total experimental time) required for all twelve judges was approx. 40 h (2327 min), requiring 109 experimental sessions.

The longer times reflect the fact that identifying even these few attributes and categorizing them as strong, weak or absent were not simple tasks. Mere preliminary presentation of standards in a descriptive analysis would not be enough to achieve this goal.

Table 2
Experimental times and numbers of experimental sessions required for each stage of training

Judge	Age (yrs)	Sex	Experimental time (min) and numbers of sessions									
			Total screening time		1st low conc. discrimination		2nd high vs. low conc. discrimination		Simultaneous identification		Serial identification	
			Time	Sessions	Time	Sessions	Time	Sessions	Time	Sessions	Time	Sessions
A	27	M	95	8	35	3	29	3	19	1	12	1
B	21	F	110	8	17	2	41	2	29	2	23	2
C	33	M	120	6	19	1	14	1	73	3	14	1
D	20	M	135	6	34	1	16	1	10	1	75	2
E	20	F	169	8	69	3	50	2	26	2	24	1
F	22	F	182	9	28	3	39	2	92	3	23	1
G	25	M	197	11	148	6	25	3	17	1	7	1
H	22	F	205	9	13	2	40	2	127	4	25	1
I	20	F	216	9	47	3	61	2	44	2	64	2
J	20	F	219	9	23	2	52	3	130	3	14	1
K	20	F	300	12	20	3	81	4	170	4	29	1
L	21	M	419	14	32	3	257	7	123	2	7	2

At this point, before beginning the descriptive analysis, the judges had been trained to identify each of the four attributes. They had also potentially established three intensity exemplars for each attribute: 'high', 'low' and 'absent'. It was hypothesized that this training was sufficient for them to achieve criterion performance for the descriptive analysis sooner with the more rapid serial monadic protocol than for the more laborious attribute-by-attribute protocol.

3. Descriptive analysis

Immediately after training (1–2 days), judges were required to perform a descriptive analysis for the four test beverages. Each had two added flavors, one with a high and one with a low concentration, with the remaining two flavors absent. Two experimental protocols were used: serial monadic and attribute-by-attribute. Testing was continued until judges could achieve criterion performance in their descriptive analysis for both paradigms. Criterion performance meant that for each attribute, the scores given to each of the four beverages accorded, at least in rank order, with their 'high', 'low' or 'absent' status. Because of the possibility of flavor intensity change due to the interaction of the added flavorings in the mixture, a third procedure was used as a control. This was to ensure that a flavoring at high concentration was perceived as more intense than the same flavoring at its lower concentration, even though other flavors were present. The control experiment consisted of sets of attribute specific discrimination tests between the four beverages.

3.1. Materials and methods

3.1.1. Stimuli

For the descriptive analysis, four test beverages were prepared, as in the training section. Each one of the four had only two added flavorings: one at a high concentration, and a different one at a low concentration (see Table 1). The arrangement of concentrations for the flavorings was balanced over the beverages (see Table 3). From the table, it can be seen that each flavoring occurred at a high concentration in one beverage, a low concentration in a second beverage, while being absent from the remaining two. For the difference testing control experiment, the stimuli to be discriminated were the test beverages (see Table 4). They were presented in approx. 10 ml aliquots in 10 ml Pyrex beakers as in the training. For the descriptive analysis, where re-tasting was more frequent, judges were presented with approx. 20 ml aliquots in 50 ml Pyrex beakers; Judges could ask for more beverages if desired.

Table 3
Identity of the stronger and weaker flavors added to each of the four beverages for the descriptive analysis experiment

	Flavorings			
	Citral	Eugenol	Green	Ionone beta
Stimulus 1	Strong	Weak	X	X
Stimulus 2	X	X	Strong	Weak
Stimulus 3	Weak	X	X	Strong
Stimulus 4	X	Strong	Weak	X

X: Absent.

3.1.2. Judges

The 12 judges who had achieved criterion performance in the training part of the study were chosen. Their details are given in Table 2.

3.1.3. Procedures

Judges were tested under three experimental protocols. They performed the descriptive analysis on the four test beverages using two experimental protocols: serial monadic and attribute-by-attribute. The third protocol was a control experiment using difference tests. Each protocol was performed in a separate experimental session on successive days, the order of testing remaining constant for a given judge. This order, however, was counterbalanced over judges. Once a judge had achieved criterion performance for one protocol, testing was ceased for that protocol but continued for the others. Feedback was given so as to enable judges to improve sufficiently to reach criterion performance.

Table 4
Stimuli used for the paired comparison (2-AFC) difference testing in the control experiment for descriptive analysis

Attribute question	Stronger flavoring	Weaker flavoring
Which has more green?	Green(H)/IB(L)[a]	Citral(H)/eugenol(L)
	Green(H)/IB(L)	Green(L)/eugenol(H)
	Green(H)/IB(L)	IB(H)/citral(L)
	Green(L)/eugenol(H)	Citral(H)/eugenol(L)
	Green(L)/eugenol(H)	Base lemon beverage
	Green(L)/eugenol(H)	IB(H)/citral(L)
Which has more citral?	Citral(H)/eugenol(L)	Green(L)/eugenol(H)
	Citral(H)/eugenol(L)	Green(H)/IB(L)
	Citral(H)/eugenol(L)	IB(H)/Citral(L)
	IB(H)/citral(L)	Green(H)/IB(L)
	IB(H)/citral(L)	Green(L)/eugenol(H)
	IB(H)/citral(L)	Base lemon beverage
Which has more eugenol?	Green(L)/eugenol(H)	IB(H)/citral(L)
	Green(L)/eugenol(H)	Green(H)/IB(L)
	Green(L)/eugenol(H)	Citral(H)/eugenol(L)
	Citral(H)/eugenol(L)	Green(H)/IB(L)
	Citral(H)/eugenol(L)	Base lemon beverage
	Citral(H)/eugenol(L)	IB(H)/citral(L)
Which has more IB?	IB(H)/citral(L)	Green(L)/eugenol(H)
	IB(H)/citral(L)	Citral(H)/eugenol(L)
	IB(H)/citral(L)	Green(H)/IB(L)
	Green(H)/IB(L)	Base lemon beverage
	Green(H)/IB(L)	Green(L)/eugenol(H)
	Green(H)/IB(L)	Citral(H)/eugenol(L)

[a] (H) and (L) refer to the flavoring being in high or low concentrations, respectively.

Before starting the serial monadic protocol, judges first sampled the nine standards used during screening (base lemon beverage, the base lemon beverage with each of the flavors at a low or high concentration). The standards were available for ad-lib re-tasting throughout a session. Judges were then presented with each of the four test beverages for assessment. The judge was required to rate each beverage on each of the four attributes, using a 10-point unstructured category scale, labeled at the ends with 'strongest' (9) and 'absent' (0). The judge was given a score sheet with four scales, one for each attribute. After assessing one test beverage, the score sheet and beverage were removed. A new score sheet, the next test beverage and standards for prior tasting were then presented for the next assessment. Judges assessed all four beverages in this manner per session. Stimuli and data from previously tested beverages were not available to judges during testing. This procedure minimized comparison of attribute scores between stimuli (attribute-by-attribute comparisons), while maintaining the serial monadic protocol. The order of presentation of the beverages was varied randomly from session to session. Judges rinsed ad-lib between re-tasting a beverage and at least three times between tasting separate beverages.

A judge was considered to have attained criterion performance if for each of the four test beverages, the scores given to the four attributes in each beverage accorded to their 'high' 'low' and 'absent' status, as described above, on two successive sessions.

For the attribute-by-attribute protocol, judges were presented with all four flavored test beverages and the base lemon beverage simultaneously for ad-lib tasting (see Table 3). The strength of a given single target flavor was rated for all five beverages. The judge was given a score sheet with five 10-point intensity scales (0–9, as above) for the relevant flavoring attribute. Re-tasting and the modification of ratings were allowed as desired. For simplicity, the judge was only presented with the three necessary standards for the attribute under consideration ('high', 'low', and 'absent'). Having rated all the stimuli on one attribute, they were all rated again in the same manner, on a second attribute, and so on. Judges rinsed ad-lib between stimuli but rinsed at least three times at the beginning of the assessment of a new attribute. At the end of testing each attribute, the score sheet and stimuli were removed, and a new score sheet, new stimuli and standards for the next attribute were presented. In each experimental session, all four attributes were used for rating all four beverages. The order of presentation of each of the four beverages was varied randomly. Judges continued to test each attribute until they achieved criterion performance as defined above.

For the control difference test sessions, judges were required to perform paired comparisons, using the four flavored test beverages prepared for the descriptive analysis. They were required to discriminate between beverages with a high concentration of a target flavor vs. beverages with a low, or zero concentration of the same flavor. They were also required to discriminate in the same way between beverages with a low level vs. a zero level of the target flavor. The task had added difficulty, because of the presence of the secondary flavors. For each flavor, there were six possible paired comparisons as illustrated in Table 4. For example, the three beverages with the higher 'green' concentration (with a low concentration of Ionone Beta). were compared with the three other beverages (one with 'low green', two with no 'green'). The next three beverages with the lower 'green' concentration ('high eugenol') were compared to the two beverages containing no green and the base lemon drink. This gave a total of six possible paired comparison for 'green'. The same procedure was used for the other flavors. Each time a target flavor was tested, the order of the six paired comparisons was varied randomly. Criterion performance required the judge to perform each given set of six paired comparisons without error for all four flavors in two successive sessions.

During testing, judges were able to taste and re-taste stimuli as often as desired. All nine standards were available for tasting throughout the testing. However, at the beginning of a series of paired comparisons, judges were required to taste the relevant standards for the flavoring under consideration. Interstimulus rinses were taken ad-lib with at least three rinses taken before testing a given attribute. All judges attained criterion performance indicating that it could be concluded that any interaction between the flavors did not prevent judges distinguishing between the three levels of each flavor (high, low, absent) in the presence of the other flavors.

An experimental session required judges to assess the four beverages using descriptive analysis or to perform six sets of difference tests. However, the stimuli were fatiguing, and some sessions had to be stretched over more than one day. Judges would continue with an experiment until they felt that the effects of 'taste fatigue' were interfering with their judgment. Generally, judges continued testing for approximately 45 min, although sometimes with difference tests, they might perform with frequent rests for periods of up to three hours. Times for the whole experiment, including the prior training, ranged 2–4 months.

At the end of every experimental session, judges were told whether any errors had been made and the offending stimuli were re-tasted. However, to avoid giving feedback that might have interfered with the learning process, the nature of the errors was not specified. Subjective reports indicated that judges were unaware that only four beverage stimuli were being used in the experiment; they believed there were many more combinations of attributes.

3.2. Results

Judges attained criterion performance for both protocols. For the serial monadic protocol, judges require 14–53 sessions ($\overline{X} = 33$); for the attribute-by-attribute protocol, 13–41 sessions ($\overline{X} = 32$). For each protocol, the number of minutes required to reach criterion performance for each judge are given in Table 5. These times included

Table 5
Time required to attain criterion performance for the serial monadic and attribute-by-attribute protocols

Judge	Attribute-by-attribute	Monadic
	Time (min)	Time (min)
A	162	39
D	165	55
L	170	165
J	199	135
C	223	121
F	266	99
I	306	84
K	386	232
G	400	181
B	406	138
H	431	203
E	1016	381
Mean*	344[a]	153[b]

* Means with different superscripts are significantly different ($p = 0.002$).

the two sessions for which criterion performance was performed. Even though, the number of sessions required to achieve criterion performance was comparable for each protocol, the mean experimental time was significantly less in the serial monadic case (t test, $p = 0.002$). All judges (binomial $p < 0.006$) followed this trend.

Error rates were compared for the initial sessions for each protocol. For the serial monadic session, there are 16 possible errors. Each flavored beverage would have one attribute at a high level (should have highest score), one at a low level (should have lower score) and two completely absent (should have zero scores). If an attribute rating did not conform to this rank order, it would be considered to be 'in error'. With four attributes, there is a total of four errors per beverage and thus sixteen per session. For each attribute-by-attribute session, inclusion of the base lemon beverage brings the total number of stimuli to be assessed per attribute to 5. This increases the total possible number per session to 20. The total percentage number of errors from the initial two sessions (8 stimuli for serial monadic; 10 for attribute-by-attribute) for the 12 judges was computed. The initial two sessions were used rather than the first session alone, to minimize any possible 'first time' effects. The percentage errors for serial monadic was less (36.7%) than for attribute-attribute (43.8%) but not significantly so (t-test, $p = 0.22$). It would appear the prior training had established sufficient intensity exemplars to attain an accuracy level comparable to the attribute-by-attribute method. This is in contrast to the results obtained with untrained judges (Ishii et al., 2007).

In the initial stage of the experiment, some judges reported that the attribute-by-attribute paradigm might be slightly easier while both protocols were very difficult and fatiguing. However, once they have gained more experience, most judges reported that they found serial monadic task easier. This observation is indeed illustrated in Table 6 as the mean experimental session lengths to complete an

Table 6
Mean experimental session lengths per judge for the initial two and last two descriptive analysis sessions, for the serial monadic and attribute-by-attribute protocols

Sessions	Attribute-by-attribute (min)	Serial monadic (min)
Initial two	7.9[a]	4.0[b]
Final two	8.2[c]	3.8[d]

[a,b,c,d] For each row, mean values with different superscripts are significantly different ($p \leqslant 0.0002$).

experimental session for the initial two sessions and final two sessions were significantly shorter for the serial monadic method than for the attribute-by-attribute method (t-test, $p \leqslant 0.0002$). This contrasts with subjective reports from untrained subjects who found the attribute-by-attribute protocol easier (Ishii et al., 2007).

4. Discussion

The hypothesis that judges who had established the appropriate number of intensity exemplars would attain criterion performance more rapidly was confirmed. Also, the fact that the serial monadic sessions ended up as being shorter in length, demonstrates their efficiency. This study and the previous study (Ishii et al., 2007) suggest that the attribute-by-attribute protocol was more suitable for naïve judges such as consumers, while after the sort of training that would be appropriate for a descriptive panel, the serial monadic protocol could be more suitable.

Judges began the descriptive analysis potentially equipped with three intensity exemplars for each attribute. These exemplars may have represented ranges of intensity rather than specific intensities, yet they were still sufficient to be operational for descriptive analysis. However, these exemplars were established for single flavors added to the lemon beverage. Some learning was still necessary for the establishment of stable exemplars in the situation where other flavors were present. It may be hypothesized that such learning was assisted by the performance of the control difference test sessions between the descriptive analysis sessions.

Part of training for descriptive analysis is to learn to identify and then consider separately the sensory attributes that are present together in the product. This can require considerable training and it is an interesting topic for research, concerning issues of mixture suppression and enhancement (Breslin, 1996; Breslin & Beauchamp, 1995, 1997; Calviño, Garcîa-Medina, & Cometto-Muñiz, 1990; McBride & Finlay, 1990; Pangborn, 1960; Schifferstein & Frijters, 1990, 1992) as well as the synthetic or analytic nature of the blending of attributes (Erickson, 1982; Erickson & Covey, 1980; Erickson, Priolo, Warwick, & Schiffman, 1990; Laing & Francis, 1989; O'Mahony, Atassi-Sheldon, Rothman, & Murphy-Ellison, 1983; Rochman, Guinard, & O'Mahony, 1997). Although these are not the issues to be studied here, it was important to check that mixture sup-

pression or synthetic blending did not prevent judges from distinguishing which flavors were 'strong', 'weak' or 'absent' during the descriptive analysis. This was established by the difference test control.

It is important to note that this study used few products and fewer attributes than would normally be encountered in a series of descriptive analyses. The study is an artificial situation made necessary by the requirement to complete the study within the limits set by the timing, personnel and stimuli. It should be regarded as preliminary and generalizations should be made with caution.

Bearing in mind these caveats about generalizability, some possibilities can still be discussed. Firstly, it is worth considering the relative model and the absolute models of cognitive processing for scaling, as two ends of a continuum rather than as a dichotomy. Naïve judges would be at the relative end of the continuum and fully calibrated judges at the absolute end. Judges who had not been fully calibrated but who had sufficient intensity exemplars to operate successfully a serial monadic descriptive panel, could be visualized as having a position on the continuum that was closer to the absolute end than the relative end. In the present study, movement along the continuum towards the absolute end, was elicited by the establishment of learned intensity exemplars during training. It is possible that such movement could also be initiated, to a certain extent, by the use of reference standards that are sometimes presented in descriptive analysis (Meilgaard et al., 1991; Stampanoni, 1993, 1994). Even, the labels on a structured category scale could be seen as moving judges away from the relative end of the continuum.

If training establishes intensity exemplars for the attributes concerned in a descriptive analysis, a question is posed. For a more typical descriptive analysis with say, 15 attributes, would training require the establishment of 15 sets of intensity exemplars or could a smaller set of exemplars be sufficient? Presumably such exemplars would be generalized over a number of attributes. It would seem to be a monumental task to establish 15 sets of intensity exemplars in memory. Perhaps only a few sets (or even one set) of flavor intensity exemplars would serve for many flavor attributes. The same might be true for texture. This is a question for further research and as such would provide useful information for designing the training of descriptive panels.

Finally, the study suggests some alternative approaches to descriptive analysis. Should a descriptive panel be required for repetitive testing, as in quality control; then it would be worthwhile spending the time on training to establish the appropriate intensity exemplars, to allow the use of the briefer serial monadic protocol. Yet, if the descriptive analysis were to be a 'one off'; then, it might be worth considering an attribute-by-attribute protocol. The descriptive analysis would take longer, but with reduced training time, the overall project time might be less.

Acknowledgements

These results were first presented to the Givaudan-Roure company in Dübendorf, Switzerland in May, 1997. The authors would like to thank Givaudan-Roure AG for their generous support and Benoit Rousseau, Miyuki Lam, Kiyoaki Okayama and Tsukasa Nakamura for their assistance.

References

Breslin, P. A. S. (1996). Interactions among salty, sour and bitter compounds. *Trends in Food Science & Technology, 7*, 390–399.

Breslin, P. A. S., & Beauchamp, G. K. (1995). Suppression of bitterness by sodium: Variation among bitter taste stimuli. *Chemical Senses, 20*, 609–623.

Breslin, P. A. S., & Beauchamp, G. K. (1997). Salt enhances flavor by suppressing bitterness. *Nature, 387*, 563.

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

Calviño, A. M., Garcîa-Medina, M. R., & Cometto-Muñiz, J. E. (1990). Interactions in caffeine–sucrose and coffee–sucrose mixtures: Evidence of taste and flavor suppression. *Chemical Senses, 15*, 505–519.

Caul, J. (1957). The profile method in flavor analysis. *Advances in Food Research, 7*, 1–40.

Dacremont, C., Sauvageot, F., & Duyen, T. H. (2000). Effect of assessors expertise level on efficiency of warm-up for triangle tests. *Journal of Sensory Studies, 15*, 151–162.

Delarue, J., & Sieffermann, J.-M. (2004). Sensory mapping using flash profile. Comparison with a conventional descriptive method for the evaluation of the flavour of fruit dairy products. *Food Quality & Preference, 15*, 383–392.

Erickson, R. P. (1982). Studies on the perception of taste: Do primaries exist?. *Physiology & Behavior 28*, 57–62.

Erickson, R. P., & Covey, E. (1980). On the singularity of taste sensations: What is a taste primary? *Physiology & Behavior, 25*, 527–533.

Erickson, R. P., Priolo, C. V., Warwick, Z. S., & Schiffman, S. S. (1990). Synthesis of tastes other than the 'primaries': Implications for neural coding theories and the concept of 'suppression'. *Chemical Senses, 15*, 495–504.

Gacula, M. C. (1997). *Descriptive sensory analysis in practice*. Trumbull, Connecticut: Food & Nutrition Press Inc.

Hootman, R. C. (1992). *Descriptive Analysis Testing*. *Manual Series* (MNL13). Philadelphia, Pennsylvania: ASTM.

Ishii, R., Chang, H. K., & O'Mahony, M. (2007). A comparison of serial monadic and attribute-by-attribute protocols for simple descriptive analysis with untrained judges. *Food Quality and Preference, 18*, 440–449.

Laing, D. G., & Francis, G. W. (1989). The capacity of humans to identify odors in mixtures. *Physiology & Behavior, 46*, 809–814.

Lawless, H. T. (1999). Descriptive analysis of odors: Reality, model or illusion? *Food Quality & Preference, 10*, 325–355.

Mata-Garcia, M., Angulo, O., & O'Mahony, M. (2007). On warm-up. *Journal of Sensory Studies, 22*, 187–193.

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

McBride, R. L., & Finlay, D. C. (1990). Perceptual integration of tertiary taste mixtures. *Perception & Psychophysics, 48*, 326–330.

Meilgaard, M., Civille, G. V., & Carr, B. T. (1991). *Sensory evaluation techniques* (2nd ed.). CRC, Boca Raton, FL: CRC Press Inc.

Mellers, B. A. (1983a). Reply to Zwislocki's views on "absolute" scaling. *Perception & Psychophysics, 34*, 405–408.

Mellers, B. A. (1983b). Evidence against "absolute" scaling. *Perception & Psychophysics, 33*, 523–526.

Neilson, A. J., Ferguson, V. B., & Kendall, D. A. (1988). Profile methods: Flavor profile and profile attribute analysis. In H. Moskowitz (Ed.).

Applied sensory analysis of foods (vol. I, pp. 21–41). Boca Raton, Florida: CRC Press.

O'Mahony, M., Atassi-Sheldon, S., Rothman, L., & Murphy-Ellison, T. (1983). Relative singularity/mixedness judgments for selected taste stimuli. *Physiology & Behavior, 31*, 749–755.

O'Mahony, M., Thieme, U., & Goldstein, L. R. (1988). The warm-up effect as a measure of increasing the discriminability of sensory difference tests. *Journal of Food Science, 53*, 1848–1850.

O'Mahony, M., & Wong, S.-Y. (1989). Time–intensity scaling with judges trained to use a calibrated scale: Adaptation, salt and umami tastes. *Journal of Sensory Studies, 3*, 217–236.

Pangborn, R. (1960). Taste interrelationships. *Food Research, 25*, 245–256.

Pfaffmann, C. (1954). Variables affecting difference tests. In D. R. Peryam, F. J. Pilgrim, & M. S. Peterson (Eds.), *Food Acceptance testing methodology. A symposium* (pp. 4–20). Washington, DC: National Academy of Sciences – National Research Council.

Rochman, D., Guinard, J.-X., & O'Mahony, M. (1997). Eliminating artifacts in the study of singularity/mixedness of taste stimuli. *Journal of Sensory Studies, 12*, 181–193.

Sieffermann, J.M. (2000). Le Profil Flash: Un outil rapide et innovant d'evaluation sensorielle descriptive. In *L'Innovation: de l'idée au succès*. Recontres Agoral 2000. Douzièmes Recontres Scientifiques et Technologiques des Industries Alimentaires, 22 et 23 Mars, Montpelier, (pp. 335–340). Paris, TEC&DOC.

Schifferstein, H. N. J., & Frijters, J. E. R. (1990). Sensory integration in citric acid/sucrose mixtures. *Chemical Senses, 15*, 87–109.

Schifferstein, H. N. J., & Frijters, J. E. R. (1992). Contextual and sequential effects on judgments of sweetness intensity. *Perception & Psychophysics, 52*, 243–255.

Sjöström, L. B., Cairncross, S. E., & Caul, J. F. (1957). Methodology of the Flavor Profile. *Food Technology, 11*, 20–24.

Stampanoni, C. R. (1993). The 'Quantitative Flavor Profiling' technique. *Perfumer & Flavorist, 18*, 19–24.

Stampanoni, C. R. (1994). The use of standardized flavor languages and quantitative flavor profiling techniques for flavored dairy products. *Journal of Sensory Studies, 9*, 383–400.

Stone, H., & Sidel, J. (1993). *Sensory evaluation practices*. Academic Press, San Diego, CA: Academic Press.

Stone, H., & Sidel, J. (1998). Quantitative descriptive analysis: Developments, applications and the future. *Food Technology, 52*(August), 48–52.

Stone, H., Sidel, J., Oliver, S., Woolsey, A., & Singleton, R. C. (1974). Sensory evaluation by quantitative descriptive analysis. *Food Technology, 28*(November), 24–34.

Thieme, U., & O'Mahony, M. (1990). Modifications to sensory difference test protocols: the warmed up paired comparison, the single standard duo-trio and the A – Not A test modified for response bias. *Journal of Sensory Studies, 5*, 159–176.

Zwislocki, J. J. (1983). Absolute and other scales: Question of validity. *Perception & Psychophysics, 33*, 593–594.

Zwislocki, J. J., & Goodman, D. A. (1980). Absolute scaling of sensory magnitudes: A validation. *Perception & Psychophysics, 28*, 28–38.