What is the validity of the sorting task for describing beers?
A study using trained and untrained assessors.

Maud Lelièvre [a,b,1], Sylvie Chollet [a], Hervé Abdi [c], Dominique Valentin [b]

[a] Institut Supérieur d'Agriculture, 59046 Lille Cedex, France

[b] UMR CSG 5170 CNRS, Inra, Université de Bourgogne, 21000 Dijon, France

[c] The University of Texas at Dallas, Richardson, TX 75083-0688, United States

ABSTRACT

In the sensory evaluation literature, it has been suggested that sorting tasks followed by a description of the groups of products can be used by consumers to describe products, but a closer look at this literature suggests that this claim needs to be evaluated. In this paper, we proposed to examine the validity of the sorting task to describe products by trained and untrained assessors. The experiment reported here consisted in two parts. In a first part, participants sorted nine commercial beers and then described each group with their own words or with a list of terms. In a second part, participants were asked to match each beer with one of their own sets of descriptors. The matching task was used to evaluate the validity of the sorting task to describe products. Results showed that 1) the categories of trained and untrained assessors were comparable, 2) trained and untrained assessors did not describe groups of beers similarly, 3) for both groups, the results of matching task were not very good and presented a high inter-variability, and 4) providing a list of terms did not seem to help the assessors. Overall, the results suggest that the sorting task followed by a description does not seem to be adapted for a precise and reliable description of complex products such as beers but may be an interesting tool to probe assessors' perception.

Key-words: Sorting task; Description; Experts; Consumers; Beer; DISTATIS, Matching task.

1[1] Corresponding author. Address: Institut Supérieur d'Agriculture, 48 Boulevard Vauban, 259046 Lille Cedex, France. Tel.: +33 3 28 38 48 01; fax: +33 3 28 38 48 47.

3E-mail address: m.lelievre@isa-lille.fr

4

INTRODUCTION

The sorting task is a simple procedure for collecting similarity data in which participants group together stimuli based on their perceived similarities. It is based on categorization which is a natural cognitive process routinely used in everyday life, and it does not require a quantitative response. This method has been routinely used by psychologists since the 1970s (e.g., Coxon, 1999; Healy and Miller, 1970). In the sensory domain, sorting tasks were first used to investigate the perceptual structure of odors (Lawless, 1989; Lawless and Glatter, 1990; MacRae, Rawcliffe, Howgate and Geelhoed, 1992; Stevens and O'Connell, 1996; Chrea et al., 2005). Lawless, Sheng and Knoops (1995) were the first to use a sorting task with a food product (cheese). Today, a large variety of products (food or non food) have been studied with this method (see Abdi, Valentin, Chollet and Chrea, 2007, for a review). Results of sorting tasks are generally analyzed using multidimensional scaling (MDS) or variation of this method (e.g., DISTATIS, Abdi et al., 2007), or sometimes with additive trees (Abdi, 1990; Corter, 1996). Generally, authors using the sorting task report that it is an easy and rapid method for obtaining perceptual maps of a large set of products, even with untrained participants.

Some authors proposed to go one step further by adding a description phase to the sorting task in order to describe the products (Lawless et al., 1995; Tang and Heymann, 1999; Saint-Eve, Paçi Kora and Martin, 2004; Faye et al., 2004, 2006; Lim and Lawless, 2005; Cartier et al., 2006; Blancher et al., 2007). So after they have sorted their products, participants are asked to describe each group with words, which are then projected onto the perceptual map of the products. Using this procedure Faye et al. (2004) studied the visual description of plastic pieces and compared the results of a free sorting task with description performed by consumers to a sensory profile performed by experts. These authors found that the conclusions reached with these two methods were quite similar for the product configurations and the words used to describe the products. Likewise, Faye et al. (2006) showed that the MDS positioning of leather samples obtained from a sorting task with description performed by consumers on visual and tactile characteristics was comparable to the sensory profile of experts. Moreover, these authors found that consumers and experts were providing related descriptions. However, these two studies involved non-food products and their results might not generalize to food products. In fact, the authors suggest that their results were specific to the case of visual and tactile senses and that their samples were easy to differentiate. In the food domain, the most recent study comparing a sorting task and a

descriptive analysis method is reported in Blancher et al. (2007). In this study, a conventional profile of visual appearance and texture of jellies was compared to a sorting task with description and a Flash profile which combined the free choice profiling and a comparative evaluation of all the products (Dairou and Sieffermann, 2002; Delarue and Sieffermann, 2004). The authors found that the Flash profile and the sorting task provided sensory maps similar to those of conventional profile for both a French and a Vietnamese panels but that the configurations obtained with the conventional profile were more similar to the configurations obtained with the Flash profile than to those obtained with the sorting task. Another recent paper from Cartier et al. (2006) showed similar results between a quantitative descriptive analysis and a sorting task with description on breakfast cereals. In this work, trained assessors performed a quantitative descriptive analysis on a set of 14 commercial breakfast cereals by rating 22 attributes of texture and flavor. Then, the same trained assessors and a group of untrained assessors performed a sorting task on the same set of breakfast cereals followed by a description of their groups of products. The authors found that products were grouped similarly in the MDS configurations derived from the sorting task and in the principal component analysis configurations derived from the sensory profile. Products were described with more terms in the sensory profile than in the sorting task and even though many terms were common to both methods, the descriptions of the groups of products were not exactly the same, especially for untrained assessors. The authors concluded that the sorting task associated with a description is a time-effective alternative to the quantitative descriptive analysis because the sorting task can provide a rough description of a large set of products. Nevertheless, some critical points emerge from a careful reading of the literature.

Several works comparing trained and untrained assessors on categorization tasks reveal that the untrained assessors' descriptions are not always comparable to the experts' descriptions. Actually, many authors report that trained assessors tend to be more efficient in their description than untrained assessors. For example Soufflet, Calonnier and Dacremont (2004) found that experts showed better abilities than untrained assessors in verbalizing their haptic perceptions of fabrics. In the food domain, Lawless et al. (1995) found that several attributes used to describe groups of cheeses were significant when regressed through the MDS space but that cheese expert assessors had a larger number of significant attributes. Saint-Eve et al. (2004)—writing about yoghourts—as well as Lim and Lawless (2005)—writing about taste solutions—found that some consensus in description was possible but all these authors also showed that untrained assessors did not agree on the verbal labeling of the groups of

products and that several of their terms were idiosyncratic. Along the same line, Piombino, Nicklaus, Le Fur, Moio and Le Quéré (2004) underlined the heterogeneity of the criteria used by assessors to characterize their groups of wines. The authors explained that among other reasons, this heterogeneity could be linked to a lack of training in the identification and description of odors. Moreover, it has been already shown with other sensory methods, such as matching or description tasks, that the attributes generated by consu are more ambiguous, redundant and less specific than the attributes generated by trained assessors (Clapperton and Piggott, 1979; Gains and Thomson, 1990; Solomon, 1990; Guerrero, Gou, and Arnau, 1997; Sokolow, 1998; Chollet and Valentin, 2001, 2006; Chollet, Valentin, and Abdi, 2005).

Another aspect never addressed in the literature is the difficulty to analyze the vocabulary used by assessors—especially consumers—to describe their groups of products. In fact, in all the studies using a sorting task, the number of terms quoted by the assessors was very large and the descriptions varied a lot from one untrained assessor to the other. Moreover, assessors spontaneously qualified their attributes with some various quantitative terms such as “very,” “many,” “slightly,” etc. So it is often necessary to preprocess the attributes before projecting them onto the MDS maps by categorizing similar terms, eliminating hedonic and idiosyncratic terms and keeping only terms cited by more than a few assessors (Cartier et al., 2006; Faye et al., 2004, 2006; Soufflet et al., 2004). This preprocessing requires time and can lead to a loss of information because it depends upon the subjectivity of the sensory analyst.

In the literature, the sorting task associated with a description performed by untrained assessors is presented as an interesting descriptive tool but is this method really valid for describing products? In order to be used for different industrial applications, the information from product descriptions has to be clearly interpretable and valid. If a description reflects the sensory properties of a given product then this product should be matched to this description. In this study, we were interested in examining the validity of the product descriptions obtained via a sorting task associated with a description. Trained and untrained assessors performed a sorting task with description followed by a matching task on nine commercial beers. The technique of matching has been already used by several authors, especially in wine domain, to evaluate expert descriptions. Lehrer (1975), followed by Lawless (1984) reported that experts were not really better in matching descriptions than untrained assessors. In contrast, Solomon (1990) found that experts clearly outperformed untrained assessors whereas Gawel (1997) showed that untrained experienced assessors

were able to outperform trained experienced assessors when they matched consensual expert descriptions. In beer domain, Chollet and Valentin (2001) found that trained and untrained assessors performed the matching task equally well, even if trained assessors were better on supplemented beers and untrained ones on commercial beers. In this study, the matching task was used to test the validity of the sorting task to describe beers as it was already done for the quantitative descriptive profile (Sauvageot and Fuentès, 2000; O'Neill, Nicklaus and Sauvageot, 2003). The validity of the sorting task was studied in a condition where assessors freely described their groups and in a condition where assessors had to choose their terms from a list (Lawless, 1988; Hughson and Boakes, 2002). By using these two conditions, we wanted to test if the use of a list of terms could help assessors, especially untrained assessors, to provide more relevant descriptions of beers.

MATERIAL AND METHODS

Assessors

Trained assessors. Thirteen assessors (5 women and 7 men) aged between 25 to 53 years (mean age = 34.9 years, SD = 9.2 years) participated. Assessors were staff members from the Catholic University of Lille (France). They had been trained one hour per week for two to five years (depending on the assessors, mean = 3.4 years, SD = 1.6 years) to detect and identify flavors (almond, banana, butter, caramel, cabbage, cheese, lilac, metallic, honey, bread, cardboard, phenol, apple, and sulfite) added in beer and to evaluate, using a non-structured linear scale, the intensity of general compounds (bitterness, astringency, sweetness, alcohol, hop, malt, fruity, floral, spicy, sparklingness, and lingering).

Untrained assessors. Two different groups of untrained assessors who were students and staff members of the University of Bourgogne (France) participated. Group A consisted of 19 assessors (6 women and 13 men) aged between 22 to 56 years (mean age = 26.6 years, SD = 8.0 years). Group B consisted in 18 assessors (19 women and 9 men) aged between 21 to 31 years (mean age = 24.6 years, SD = 2.4 years). They were beer consumers but did not have any formal training or experience in the description of beers.

Products

Nine different commercial beers were evaluated (denoted PelfBL, PelfA, PelfBR, ChtiBL, ChtiA, ChtiBR, LeffBL, LeffA and LeffBR). These beers came from three different breweries: *Pelforth* (noted Pelf), *Chti* (Chti) and *Leffe* (Leff) and each brewery provided

three types of beer: blond (BL), amber (A) and dark (BR). All beers were presented in three-digit coded black plastic tumblers and served at 10°C.

Procedures

Experiment

Subjects took part individually in the experiment in a single session. The experiment was conducted in separate booths lighted with a neon lighting of 18 watts with a red filter darkened with black tissue paper to mask the color differences between beers. Mineral water and bread were available for assessors to rinse between samples. Assessors could spit out beers if they wanted.

The experiment consisted in two parts. The first one was a sorting task and the second a matching task. These two parts are explained below.

Part 1. Sorting task with description: The assessors received the entire set of beers. The order of presentation of the samples was performed according to a Latin Square. Panelists were first required to smell and taste each sample once in the proposed order. Afterward, they were allowed to smell and taste samples as many times as they wanted and in any order. No criterion was provided to perform the sorting task. Assessors were free to make as many groups as they wanted and to put as many beers as they wanted in each group. They were allowed to take as much time as they wanted. After they had finished their sorting task, the assessors were asked to describe each group of beers with some words according to two conditions. In the first condition, assessors were free to use their own words. In the second condition, assessors had to choose their words from a list of 44 terms which were extracted from the Flavor Wheel of the International Terminology System for Beer (Meilgaard, Dalgliesh, and Clapperton, 1979).

Table 1. List of the 44 terms used for the second condition (from Meilgaard et al., 1979)

1. Alcoholic	23. Sulfidic
2. Solvent like	24. Cooked Vegetable
3. Estery	25. Yeast
4. Fruity	26. Stale
5. Acetaldehyde	27. Catty
6. Floral	28. Papery
7. Hoppy	29. Leathery
8. Resinous	30. Moldy

9. Nutty	31. Acidic
10. Grassy	32. Acetic
11. Grainy	33. Sour
12. Malty	34. Sweet
13. Worty	35. Salty
14. Caramel	36. Bitter
15. Burnt	37. Alkaline
16. Phenolic	38. Mouthcoating
17. Fatty acid	39. Metallic
18. Diacetyl	40. Astringent
19. Rancid	41. Powdery
20. Oily	42. Carbonation
21. Sulfury	43. Warming
22. Sulfitic	44. Body

Because we had only one group of trained assessors, we used a within-subject design (all trained assessors performed the experiment in the two conditions without and with the list of terms) whereas for untrained assessors, we used a between-subject design (group A performed the task in the condition without the list and group B in the condition with the list). In both conditions (without and with the list), assessors were told to use no more than five words per group of beers and to indicate the *intensity* of the descriptors using a four-point scale labeled: “not,” “a little,” “medium” and “very.” Assessors did not know that they would have to describe their beer groups when they performed the sorting task. Also, they could not change the beer groups they had just made.

Part 2. Matching task: After a twenty-minute break, assessors received the nine beers again and were provided with the sets of terms they had just used to describe their beer groups. They were not informed that the beers were the same that the ones used for the sorting task. They were asked to match each beer with a set of terms. The instructions indicated that one beer could be associated with only one set of descriptive terms and that assessors were not obliged to use all the sets of terms (some sets of terms could be associated with no beer). When they performed the sorting task, assessors did not know that they would have to match their descriptions later on.

Data analysis

Sensory map of the products

For each assessor, the results of the sorting task were encoded in an individual distance matrix where the rows and the columns are beers and where a value of 0 between a row and

a column indicated that the assessor put the beers together, whereas a value of 1 indicated that the beers were not put together. For each group of assessors (trained and untrained group A and B) and each condition (without and with the list), the individual distance matrices obtained from the sorting data were analyzed by using DISTATIS (Abdi, Valentin, O'Toole and Edelman, 2005; Abdi et al., 2007). This method is a generalization of classical multidimensional scaling. DISTATIS takes into account individual sorting data and it provides a compromise map for the products which is a MDS-like map. This product map is obtained from a principal component analysis performed on the DISTATIS compromise cross-products matrix which is a weighted average of the cross-products matrices associated with the individual distance matrices derived from the sorting data (Abdi et al., 2007). In this map, the proximity between two points reflects their similarity. We also computed R_v coefficients between trained and untrained assessors' configurations in the two conditions with and without list. The R_v coefficient measures the similarity between two configurations and can be interpreted in a manner analogous to a squared correlation coefficient (Abdi, 2007).

Analysis of the vocabulary

Each assessor described each group of beers with words. For each assessor, the terms given for a group of products were associated to each beer of the group. We assumed that all the beers belonging to the same group were described by the terms in the same way. We began by regrouping the synonyms. Then we converted each intensity word into a score in order to obtain an intensity score for each term quoted to describe the groups of beers: "not"= 0, "a little"= 1, "medium"= 2 and "very"= 3. Then, in order to analyze the vocabulary used by trained and untrained assessors, we computed the geometric mean for each quoted term and each beer for trained and untrained assessors as described in Dravnieks (1982):

$$M = \sqrt{F \times I}$$

where F is the frequency of quotation of each term and is calculated by dividing the number of times when the term was quoted with an intensity different from zero by the maximum number of quotations for a term (number of assessors); I is the intensity for each quoted term and is computed as the sum of the intensities for the term divided by the maximal intensity for a term (number of assessors by maximum score for a term). The geometric mean is expressed as a percentage. Only terms having a geometric mean higher or equal to 20% for at least one product were considered. The geometric means of these terms were then projected onto the compromise spaces for trained and untrained assessors in the two

conditions (without and with the list), according to the method described in Abdi et al. (2007).

Evaluation of the validity of the vocabulary

To study the validity of the vocabulary used by trained and untrained assessors to describe their groups of beers, we examined the results of the matching task. We assumed that if assessors were able to make the same groups of beers from their descriptions as they did during the sorting task, then the terms they used to describe their groups of beers were valid. We computed the number of correct matches, which corresponds to the number of times a beer was matched with the right description written during the sorting task. For convenience, the results are expressed as the percentage of correct matches. We computed Student *t*-tests between the means of the percentages of correct matches for the assessors and the means of the percentages of correct matches expected by chance. The percentage of correct matches to be expected by chance was different for each assessor because the number of descriptions differed from one assessor to another, depending on the number of sorting groups. This percentage for an assessor was computed as: (1/number of descriptions of the assessor) $\times$ 100. In order to study the effect of training (trained/untrained) and the use of a list of terms (without/with the list) on the validity of the vocabulary, Student *t*-tests were also performed on the means of the percentages of correct matches. Differences are considered significant at *alpha* = 0.05 level.

RESULTS

Figure 1 shows the compromise maps obtained for trained and untrained assessors' sorting results. Terms (only the ones with a geometric mean higher or equal to 20%) are plotted onto these maps for the two conditions without and with the list.

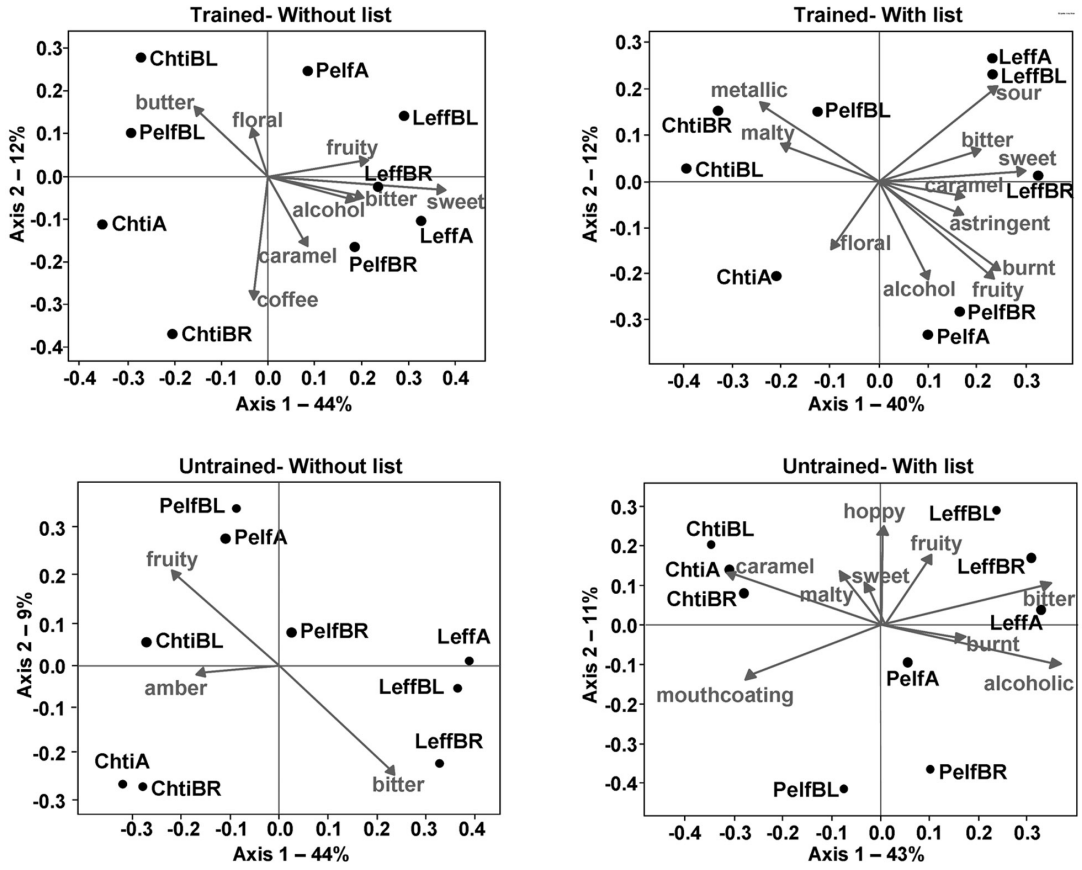

Figure 1. Two dimensional compromise maps for trained assessors (top panel) and untrained assessors (bottom panel) for their sorting tasks followed by descriptions without the list (on the left) and with the list (on the right). The geometric means of each term are plotted onto the compromise spaces.

How did trained and untrained assessors categorize beers?

As shown in Figure 1, on the whole, trained and untrained assessors categorized the nine beers in the same way. These observations were confirmed by the large values of R_v coefficients computed between trained and untrained assessors' configurations which were significant for the two conditions without ($R_v = 0.71$, $p < 0.05$) and with the list of terms ($R_v = 0.65$, $p < 0.05$). There is a clear separation of the beers into breweries. The three Chti beers are opposed to the three Leffe beers on the first dimension which explained 44% of the total variance. The three Pelforth beers are a little less well clustered. They are spread between the Chti and the Leffe beers on the first axis. They are opposed to the Chti and Leffe beers on the second dimension for untrained assessors and are more mixed with the two other breweries for trained assessors. However these differences between trained and untrained assessors for the Pelforth beers should be interpreted with caution since axis 2

only explains a relatively small amount of total variance (12% for trained and 9% for untrained assessors).

How did trained and untrained assessors describe the groups of beers?

Expertise level effect. Without any list of terms, we clearly observe a larger number of descriptors with a geometric mean above 20% for trained assessors: there were only three terms out of 54 with a geometric mean higher than 20% for untrained assessors, while there were eight out of 35 for trained assessors. The terms *fruity* and *bitter* were common to the descriptions of the two groups of assessors but only *bitter* was used to describe the same beers (Leffe beers). Globally, the descriptions of the groups of beers were different for trained and untrained assessors without the list. In the condition with the list, the number of descriptors was quite similar for trained (10 terms out of 27) and untrained assessors (9 terms out of 34) and seven terms were common to their descriptions (*malty*, *sweet*, *burnt*, *bitter*, *caramel*, *alcoholic* and *fruity*). Only *bitter* (for the three Leffe beers) and *fruity* (for LeffBL) were used to describe the same beers for the two groups of assessors.

List effect. If we compare the two conditions without and with the list for trained assessors, we find some common points: the terms *alcohol*, *sweet*, *bitter*, *caramel*, *floral* and *fruity* were common to both descriptions. In the two conditions, trained assessors described Leffe beers as *sweet*, *fruity*, *bitter* and *caramel*. However, we can note some differences. For example, trained assessors characterized ChtiBL with the term *butter* only in the condition without the list. Also, they described PelfA with *floral* without the list and with *astringent* and *alcohol* with the list. Along the same line, ChtiBR was characterized using the attribute *coffee* without the list and as *metallic* and *malt* with the list. Concerning untrained assessors, we observe that they used many more terms with the list than without the list. For example with the list, they described beers with terms such as *hop*, *malt*, *caramel*, *alcoholic*, *burnt*, *sweet*, or *smooth*. Two terms were common to the two descriptions without and with the list: *bitter* and *fruity*, but only *bitter* characterized the same beers in the two conditions (Leffe beers). Moreover, a more detailed analysis of the raw data shows that the terms *hop* and *malt* were used by untrained assessors to describe all of the nine beers whereas trained assessors never used *hop* to describe the beers and *malt* was only used for ChtiBL.

Quantitative terms. We examined how trained and untrained assessors used the four quantitative words: “not”, “a little”, “medium” and “very”. We found that trained assessors

used the words “very” twice as often as “a little.” In contrast, untrained assessors used the three terms “a little,” “medium” and “very” in a similar way. Moreover, untrained assessors used the word “not” to characterize their descriptors more frequently (20 times) than trained assessors (5 times) did ($\chi^2 = 9$, d.f. = 1, $p < 0.01$).

What is the validity of the terms used by trained and untrained assessors?

Student *t*-tests showed that the results of trained assessors were significantly better than chance when assessors matched their descriptions for the two conditions [Average(without the list) = 54.7%, $t(12) = 2.82$, $p < 0.01$; Average(with the list) = 59.0%, $t(12) = 4.39$, $p < 0.001$], as well as the results of untrained assessors [Average(without the list, group A) = 50.9%, $t(18) = 4.49$, $p < 0.001$; Average(with the list, group B) = 48.1%, $t(17) = 4.10$, $p < 0.001$].

Student *t*-tests did not detect a difference between the two conditions without and with the list for trained assessors ($t(12) = 0.50$, ns), and for untrained assessors ($t(35) = 0.36$, ns). In the same way, there was no statistically significant difference between the two groups of assessors in the condition without the list ($t(30) = 0.36$, ns) as well as in the condition with the list ($t(29) = 1.28$, ns). So there was no statistically significant difference on the validity of the vocabulary neither between trained and untrained assessors nor between the two conditions (without/with the list). However, this failure to show any significant effect can be explained by the large inter-individual variability of the results.

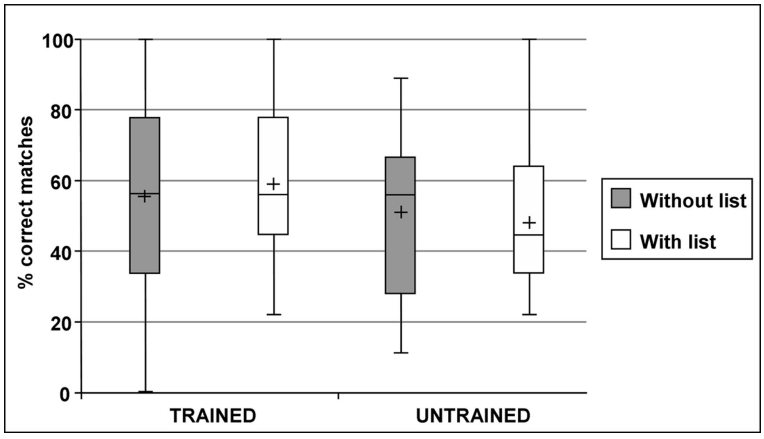

Figure 2. Box plot of percentage of correct matches distributions calculated for trained and untrained assessors in the two conditions without (black boxes) and with the list (white boxes), for the matching task.

Figure 2 shows the box plot of the distributions of the percentage of correct matches for trained and untrained assessors in the two conditions (without and with the list). The box extends from the first to the third quartile, the line across the box represents the median, the

plus sign represents the mean value and the ends of the lines extending from the box ("whiskers") indicate the maximum and the minimum data values, unless outliers are present in which case the whiskers extend to a maximum of 1.5 times the inter-quartile range (i.e. length of the box). In our case, the whiskers represent the extreme values. We can see a high inter-individual variability especially for trained assessors in the condition without the list. A finer grained analysis of the raw data shows that three trained assessors perfectly succeeded in the matching task (percentage of correct matches = 100%) and two trained assessors did not succeed at all in associating the beers with their descriptions (percentage of correct matches = 0%).

DISCUSSION

In recent years, using sorting tasks associated with a description with consumers has started to become a popular way of describing food and non-food products. This approach proved to be useful to obtain a coarse description of products (Blancher et al., 2007; Cartier et al., 2006; Faye et al., 2004, 2006; Tang and Heymann, 1999; Saint-Eve et al., 2004) but can it be considered as a plausible alternative to conventional profiling? The information conveyed by products descriptions has numerous applications in product development, quality control or consumer preference understanding. Thus, because of these important and widespread applications, the information conveyed by products descriptions needs to be clearly interpretable, reliable and valid. To this extent, a product description should convey the sensory properties of the product it represents in such a way that a product can be matched to its corresponding description. In this study, we examined if product descriptions obtained via a sorting task associated with a description could match this requirement. We compared the performance of trained and untrained assessors in two description conditions (without and with a list of terms).

Are trained and untrained assessors comparable?

To address this question, we compared trained and untrained assessors descriptions. In the condition without list, we found that the descriptions of the groups of beers were rather different for both groups of assessors. This result does not replicate Cartier et al.'s (2006) study which found that the descriptions of groups of breakfast cereals were almost similar between trained and untrained assessors. We observed that there were many more terms quoted by untrained assessors (54 terms) than by trained assessors (35 terms). But when selecting only terms with a geometric mean above 20%, only three terms for untrained

assessors and eight terms for trained assessors were kept. This result reflects the lack of consensus in both the choice of the terms and in perceived intensity, especially for untrained assessors. The greater lack of agreement among untrained assessors in comparison to the trained assessors is not very surprising. Indeed, training involves the development of a common lexicon with standard physical references allowing an alignment and a standardization of the sensory concepts of the panelists (Ishii and O'Mahony, 1990). The importance of training in reaching a consensus is illustrated by the fact that seven out of the eight terms of trained assessors were attributes belonging to the profile list of attributes used for their training. For example, a trained assessor described the three Leffe beers in this way: "*very* sweet, *very* alcohol, *medium* hop, *medium* bitter," whereas an untrained assessor described these same beers with: "*medium* exotic feel, *medium* spicy sensation, *medium* grapping taste (goût prenant)." This difference between the descriptors used by trained and untrained assessors can be explained by the training of trained assessors which allows them to possess a specific and precise vocabulary. Finally we found that trained and untrained assessors used the four intensity words differently. Contrary to untrained assessors who used the three expressions "a little," "medium," and "very" in the same way, we observed that trained assessors used "very" twice as often as "a little." We also noticed that untrained assessors used the word "not" frequently, while trained assessors hardly used it. So it seems that trained assessors tend to describe their groups of beers with distinctive characteristics (i.e., characteristics with a high intensity) whereas untrained assessors do not use particular characteristics to describe their groups of beers. These observations highlight the interest of using intensity scores to quantify attributes. These quantitative words bring additional information to the descriptions and we think that it is important to impose their use to the assessors.

The comparison between trained and untrained assessors' descriptions confirmed the conclusions of several authors that trained assessors used more specific terms, especially terms learned during training (Clapperton and Piggott, 1979; Chollet and Valentin, 2001; Chollet et al., 2005). We expected this high specificity of trained assessors' vocabulary to lead to a better matching performance than that of untrained assessors. Yet, contrary to previous work (Lawless, 1984; Solomon, 1990; Gawel, 1997) we did not find any difference in matching performance between the two groups of assessors. Both trained and untrained assessors were above chance level but their performance levels were not very high (54.7% of correct matches for trained assessors and 50.9% for untrained ones). The overall low performance of trained assessors, however, might be due to the high inter-individual

variability. Indeed, while three trained assessors performed perfectly, two others were below chance level. A plausible explanation for this high variability is the difference in years of training of the panelists. Indeed, the panelists with 100% of correct matches were among the panelists who had the longest training. Yet correlation coefficient computed between the percentage of correct matches and the years of training shows that it is not the only explanation (r = .61, r^2 = 0.37, p < .05). The fact that some trained assessors with four or five years of training succeeded in the matching task whereas others had poor results may suggest that some trained assessors are better than others to generalize their knowledge to a new task. It has been already showed that trained assessors were not able to generalize their perceptual knowledge to new beers (Chollet et al., 2005). The same problem could exist with new tasks and this might be related to the duration of training.

Is providing a list helpful?

We found that the descriptions of the beers were different when assessors had a list of terms and when they did not have such a list, especially for untrained assessors. For untrained assessors, we observed a larger number of descriptors with a geometric mean above 20% with the list than without the list. This suggests that having a list of terms can be helpful for untrained assessors. But a deeper look at the descriptions with the list shows that, for example, untrained assessors used *hop* and *malt* to describe almost all the beers. It is probable that the list given to untrained assessors influenced their descriptions. The untrained assessors probably knew that *hop* and *malt* are terms associated with the brewing process and so they used it but without knowing exactly what these terms mean. We assume that the descriptions containing these words *hop* and *malt* did not allow them correct matches. For trained assessors, the number of descriptors with a geometric mean above 20% was quite similar between the two conditions. Moreover, the results of the matching task were not better with the list than without the list for both trained and untrained assessors.

The efficiency of the list in this study can be put in perspective with the results of Hughson and Boakes (2002). In this study, assessors had to describe five white wines according to three conditions: without any list of terms, with a long list of terms (125 terms) and with five short lists of terms (14 terms in each list corresponding to each wine). Then, they had to match their own descriptions to the wines. Matching performance was better in the short-list condition (40% of correct matches) than in the long-list condition (27 % of correct matches) and in the control condition without any list (16% of correct matches). Moreover, only results in the short-list condition were above chance. So we can wonder why our list did not

help assessors to improve their scores of matching too. One reason could be that our list of terms was too long (44 terms) compared to the one of Hughson and Boakes (14 terms) to help assessors to effectively describe the beers. In the case of trained assessors, another reason could be that the terms provided were different from the terms used in training. This hypothesis is supported by the fact that trained assessors described ChtiBL as *butter* in the condition without the list but did not in the condition with the list. Interestingly, in Meilgaard's list, *butter* is replaced by *diacetyl*, which is associated with the butter flavor and so trained assessors did not seem to know the term *diacetyl*. This remark highlights the importance of using a common descriptive vocabulary. Some authors such as Rainey (1986), Civille and Lawless (1986) or Stampanoni (1994) indicated that for sensory profiles, the use of a common terminology based on references reduced the time for training and improved the agreement between the assessors. In our case, the use of a terminology without associated reference did not help assessors to describe the beers. Finally, the fact that the list of terms did not help the assessors could be due to the use of a previoulsy published list which was not exactly adapted to our products. In the study of Hughson and Boakes (2002), the short lists provided to the assessors contained terms which corresponded exactly to the wines to be described.

CONCLUSION

Our results highlight some important problems that might be encountered when using a sorting task to describe a set of products, especially with untrained assessors: difficulties for analyzing the vocabulary (many terms to preprocess), high inter-individual variability, lack of precision of the descriptions and sensitivity of the used methodology (presence of a list or not). Because different descriptions are obtained depending on the experience level of assessors and the specific procedures used (with or without a list), we would suggest that sorting tasks followed by a description task provide an interesting tool to understand how assessors perceive a set of products. Thus, this method might be recommended in studies focusing on assessors' behavior. However, in order to describe precisely and reliably complex products such as beers, a training phase might be necessary and a method such as conventional profiling is probably more adapted.

ACKNOWLEDGEMENT

This work was financed by the Institut Supérieur d'Agriculture. The authors would also like to gratefully thank the anonymous reviewers who for their helpful comments on a previous version of this paper.

REFERENCES

Abdi, H. (1990). Additive-tree representations. *Lecture Notes in Biomathematics*, *84*, 43-59.

Abdi, H. (2007). The R_v coefficient and the congruence coefficient. In N. Salkind (Ed.): *Encyclopedia of measurement and statistics.* Thousand Oaks (CA): Sage. pp 849-853.

Abdi, H., Valentin, D., Chollet, S., and Chrea, C. (2007). Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. *Food Quality and Preference*, *18*, 627-640.

Abdi, H., Valentin, D., O'Toole, A.J., and Edelman, B. (2005). DISTATIS: The analysis of multiple distance matrices. *Proceedings of the IEEE Computer Society: International Conference on Computer Vision and Pattern Recognition.* (San Diego, CA, USA). pp. 42-47.

Blancher, G., Chollet, S., Kesteloot, R., Nguyen Hoang, D., Cuvelier, G., and Sieffermann, J.-M. (2007). French and Vietnamese: How do they describe texture characteristics of the same food? A case study with jellies. *Food Quality and Preference*, *18*, 560-575.

Cartier, R., Rytz, A., Lecomte, A., Poblete, F., Krystlik, J., Belin, E., and Martin, N. (2006). Sorting procedure as an alternative to quantitative descriptive analysis to obtain a product sensory map. *Food Quality and Preference*, *17*, 562-571.

Chollet, C., and Valentin, D. (2001). Impact of training on beer flavor perception and description: Are trained and untrained subjects really different? *Journal of Sensory Studies*, *16*, 601-618.

Chollet, S., and Valentin, D. (2006). Impact of training on beer flavour perception. *Cerevisia, Belgian Journal of Brewing and Biotechnology*, *31*, 189-195.

Chollet, S., Valentin, D., and Abdi, H. (2005). Do trained assessors generalize their knowledge to new stimuli? *Food Quality and Preference*, *16*, 13-23.

Chrea, C. Valentin, D., Sulmont-Rossé, C., Ly, M.H., Nguyen, D., and Abdi, H. (2005). Semantic, typicality and odor representation: A cross-cultural study. *Chemical Senses, 30*, 37-49.

Civille, G. V., and Lawless, H. T. (1986). The importance of language in describing perceptions. *Journal of Sensory Studies, 1*, 203-215.

Clapperton, J. F., and Piggott, J. R. (1979). Flavour characterization by trained and untrained assessors. *Journal of the Institute of Brewing, 85*, 275-277.

Corter, I.E. (1996). *Tree models of similarity and association*. Thousand Oaks: Sage.

Coxon, A. P. M. (1999). *Sorting data: Collection and analysis*. Thousand Oaks: Sage.

Dairou, V., and Sieffermann, J.-M. (2002). A comparison of 14 jams characterized by conventional profile and a quick original method, the flash profile. *Journal of Food Science, 67*, 826-834.

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

Dravieks, A. (1982). Odor quality: semantically generated multidimensional profiles are stable. *Science, 218*, 799-801.

Faye, P., Brémaud, D., Durand Daubin, M., Courcoux, P., Giboreau, A., and Nicod, H. (2004). Perceptive free sorting verbalization tasks with naive subjects: an alternative to descriptive mappings. *Food Quality and Preference, 15*, 781-791.

Faye, P., Brémaud, D., Teillet, E., Courcoux, P., Giboreau, A., and Nicod, H. (2006). An alternative to external preference mapping based on consumer perceptive mapping. *Food Quality and Preference, 17*, 604-614.

Gains, N., and Thomson, D. M. H. (1990). Sensory profiling of canned lager beers using novices in their own homes. *Food Quality and Preference, 2*, 39-47.

Gawel, R. (1997). The use of language by trained and untrained experienced wine tasters. *Journal of Sensory Studies, 12*, 267-284.

Guerrero, L., Gou, P., and Arnau, J. (1997). Descriptive analysis of toasted almond: A comparison between experts and semi-trained assessors. *Journal of Sensory Studies, 12*, 39-54.

Healy, A., and Miller, G. A. (1970). The verb as the main determinant of the sentence meaning. *Psychonomic Science, 20*, 372.

Hughson, A. L., and Boakes, R. A. (2002). The knowing nose: The role of knowledge in wine expertise. *Food Quality and Preference, 13*, 463-472.

Ishii, R., and O'Mahony, M. (1990). Group taste concept measurement: verbal and physical definition of the umami taste concept for Japanese and Americans. *Journal of Sensory Studies, 4*, 215-227.

Lawless, H.T. (1984). Flavor description of white wines by "expert" and nonexpert wine novices. *Journal of Food Science, 49*, 120-123.

Lawless, H.T. (1988). Odor description and odor classification revisited. In D. Thompson (Ed.), *Food acceptability*. London and New York: Elsevier Applied Science.

Lawless, H.T. (1989). Exploration of fragrances categories and ambiguous odors using multidimensional scaling and cluster analysis. *Chemical Senses, 14*, 349-360.

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

Lawless, H.T., Sheng, N., and Knoops, S.S.C.P. (1995). Multidimensional scaling of sorting data applied to cheese perception. *Food Quality and Preference, 6*, 91-98.

Lehrer, A. (1975). Talking about wine. *Language, 51*, 901-923.

Lim, J., and Lawless, H.T. (2005). Qualitative differences of divalent salts: multidimensional scaling and cluster analysis. *Chemical Senses, 30*, 719-726.

MacRae, A.W., Rawcliffe, T., Howgate, P., and Geelhoed, E.N. (1992) Patterns of odour similarity among carbonyls and their mixtures. *Chemical Senses, 17*, 119-125.

Meilgaard, M.C., Dalgliesh, C. E., and Clapperton, J. F. (1979). Beer flavor terminology. *Journal of the American Society of Brewing Chemists, 37*, 47-52.

O'Neill, L., Nicklaus, S., and Sauvageot, F. (2003). A matching task as a potential technique for descriptive profile validation. *Food Quality and Preference, 14*, 539-547.

Piombino, P., Nicklaus, S., Le Fur, Y., Moio, L., and Le Quéré, J.-L. (2004). Selection of products presenting given flavor characteristics: an application to wine. *American Journal of Enology and Viticulture, 55*, 27-34.

Rainey, B.A. (1986). Importance of reference standards in training panelists. *Journal of Sensory Studies, 1*, 149-154.

Saint-Eve, A., Paçi Kora, E., and Martin, N. (2004). Impact of the olfactory quality and chemical complexity of the flavouring agent on the texture of low fat stirred yogurts assessed by three different sensory methodologies. *Food Quality and Preference, 15*, 655-668.

Sauvageot, F., and Fuentès, P. (2000). Une approche pour valider la technique du profil sensoriel: la technique de l'appariement, *Sciences de l'Aliment, 20*, 467-489.

Sokolow, H. (1998). Quantitative methods for language development. In H. Moskowitz (Ed.), *Applied Sensory Analysis of Food*. Boca-Raton, Florida (pp. 3-19).

609 Solomon, G.E.A. (1990). Psychology of novice and experts wine talk. *American Journal of*
610 *Psychology, 103*, 495-517.
611
612 Soufflet, I., Calonnier, M., and Dacremont, C. (2004). A comparison between industrial
613 experts' and novices' haptic perception organization: a tool to identify descriptors of handle
614 of fabrics. *Food Quality and Preference, 15*, 689-699.
615
616 Stampanoni, C.R. (1994). The use of standardized flavor languages and quantitative flavor
617 profiling technique for flavored dairy products. *Journal of Sensory Studies*, *9*, 383-400.
618
619 Stevens, D.A., and O'Connell, R.J. (1996). Semantic-free scaling of odor quality.
620 *Physiological Behavior, 60*, 211-215.
621
622 Tang, C., and Heymann, H. (1999). Multidimensional sorting, similarity scaling and free
623 choice profiling of grape jellies. *Journal of Sensory Studies*, *17*, 493-509.