Available online at www.sciencedirect.com

ScienceDirect

Food Quality and Preference 18 (2007) 675–680

Food Quality and Preference

www.elsevier.com/locate/foodqual

Sensory difference tests for margarine: A comparison of R-Indices derived from ranking and A-Not A methods considering response bias and cognitive strategies

H.S. Lee[a,*], D. van Hout[a], M. O'Mahony[b]

[a] *Unilever R&D Vlaardingen, Oliver van Noortlaan 120 3133 AT Vlaardingen, The Netherlands*
[b] *Department of Food Science and Technology, University of California, Davis, CA 95616, USA*

Received 23 June 2006; received in revised form 16 October 2006; accepted 24 October 2006
Available online 8 December 2006

Abstract

Sensory difference tests were performed between 6 margarine products, a standard vs 5 other products. Three testing protocols were used. The first protocol was simple ranking. The second protocol was the A-Not A method where a single standard was presented beforehand and which could be retasted during testing. The third protocol was the A-Not A method where all products were presented beforehand but could not be retasted during testing. R-Index values were computed for each protocol. Ranking gave the highest R-Index values while the A-Not A method, where only a single standard was presented prior to testing, gave the lowest R-Index values. R-Indices were calculated by averaging indices from individual judges and also by pooling data from all judges. Differences between these computations only occurred for the A-Not A method where all the products were presented prior to testing. Differences were explained in terms of the forced-choice nature of ranking, boundary variance, concept formation and differences in cognitive strategies involving tau and beta-criteria.

Keywords: Ranking; A-Not A; R-Index; Tau and beta-criteria; Boundary variance; Cognitive strategies; Familiarization; Concept formation

1. Introduction

Sensory difference tests are used for discriminating between two confusable food stimuli or other products with sensory attributes. Such tests are used for reformulation, quality control, product development, ingredient specification, shelf-life, cost reduction, packaging studies etc. One such test is the A-Not A method, sometimes called the single stimulus method. Although A-Not A test is not the most common test, it is used by food industry. This method was first introduced to food science by Pfaffmann, Schlosberg, and Cornsweet (1954). For this protocol, a product (A) is presented to the judge several times at the start of an experimental session so that the judge can become familiar with it. Then, a series of two products, 'A' and a slightly different product to be discriminated from 'A' (Not A) are presented in random order. The judge has to respond by stating which products are 'A' and which are 'Not A'. During the test session, the judge is given 'A' at various intervals, knowing its identity, as a reminder.

Peryam (1958) later described the test in the same way except that he stated that not only 'A' but sometimes 'Not A' could also be presented at the beginning of the test for familiarization. Although both products were removed before testing, 'A' could be presented to the judge during the experimental session as a reminder, as described by Pfaffmann.

[*] Corresponding author. Present address: Department of Food Science and Technology, Ewha Womans University, Seoul 120-750, South Korea. Tel.: +82 2 3277 3624; fax: +82 2 3277 3095.
E-mail address: hlee@ewha.ac.kr (H.S. Lee).

0950-3293/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.foodqual.2006.10.003

Meilgaard, Civille, and Carr (1991) also described the test with both products being presented beforehand for familiarization. However, in their version of test, no reminders are given. They did mention that although usually only one 'Not A' product is presented during testing, sometimes more than one 'Not A' product could be presented. In such a case, all possible 'Not A' products should be presented prior to the test. ASTM (1968) described the test indicating that the two products (A and Not A) are given beforehand; they are vague on further details. Lawless and Heymann (1996) point out that there have been various versions of the A-Not A method. It would appear that there is no agreed standard A-Not A method; as long as the general procedure is followed, the method is given its name. Because of this, it is always better to describe in detail the methods being used. The different methods have the potential to change the cognitive strategy being used. If there were changes in cognitive strategy, comparisons of the discrimination indices between methods would be problematical. Because of this it is worth gaining understanding of these effects.

The problem with the A-Not A method is that like the same-different method, it has inherent response bias (O'Mahony & Rousseau, 2002). Specifically, for the A-Not A method, a judge's response will not only depend on whether his sensory system is sufficiently sensitive to discriminate between the products 'A' and 'Not A', but also depend on his willingness to report a different product as 'Not A'.

Because of such response bias, merely counting the proportion of correct responses in a A-Not A method is a biased measure of perceived difference. However, response bias can be circumvented by using a signal detection/Thurstonian analysis to calculate an index of difference such as d′ (Macmillan & Creelman, 2005; O'Mahony & Rousseau, 2002). An R-Index computation could also be used (Delwiche & O'Mahony, 1996). For such computations, judges would generally add sureness judgments to their 'A' versus 'Not A' judgments (Brown, 1974).

For the computation of d′, it is necessary to know the cognitive strategy being used in the A-Not A test (O'Mahony & Rousseau, 2002; Lee & O'Mahony, 2004; O'Mahony, Masuoka, & Ishii, 1994). There are two logical possibilities. Firstly, the A-Not A method could be treated as a version of the Yes-No procedure in signal detection which implies the use of a beta-criterion (Green & Swets, 1966). For this it is assumed that there are only two products in the test. In this case, the judge would hold two categories in his head, one corresponding to 'A' and the other to 'Not A'. The boundary between the two categories would be the beta-criterion. The judge would assign the test products to either category and respond accordingly. Recent neuropsychological research has reported that the A-Not A method using multiple stimuli (more than one type of 'Not A') also utilizes a beta-criterion (categorization) (Casale, Ashby, & Standring, 2005).

Secondly, the A-Not A method could be considered as an extension of the same-different method. This method is assumed to use a tau-criterion., although there are a few exceptions (Lee, van Hout, Hutaus, & O'Mahony, forthcoming; Irwin & Francis, 1995; Francis & Irwin, 1995), The tau-criterion is a degree to which two stimuli must differ, to be reported to be different. If the two products are more different than the tau-criterion, they are reported as different. However, if they are not more different than the tau-criterion, they are reported as the same. Accordingly, if the product to be tested is perceived as more different from 'A' than the tau-criterion, the judge will report it as 'Not-A'. If it is not perceived as more different, it will be reported as 'A'.

Should the A-Not A method be used in a situation where there are several 'Not A' products, an alternative procedure might be to use a simple ranking method. Judges would rank the products in terms of their similarity to 'A'. Designating 'A' as the 'noise', the degree of difference between 'A' and the various test products (signals) could be computed using an R-Index analysis (Brown, 1974; O'Mahony, 1992). The A-Not A method is essentially a rating or categorization procedure from which R-Index values can be computed to represent the degree of difference between 'A' and the various 'Not A' products. The ranking procedure would also give R-Index values representing the difference between 'A' and the various 'Not A' products. However, because of its forced choice nature, ranking tends to give higher R-Index values than those calculated from rating or categorization (O'Mahony, Garske, & Klapman, 1980; Ishii, Vié, & O'Mahony, 1992).

Prior research indicating that ranking gives higher R-Index values than a simple rating or categorization procedure was performed using a simple model system (O'Mahony et al., 1980; Ishii et al., 1992). The goal of the present experiment was to investigate whether ranking gave higher R-Index values than a more complex rating or categorization procedure, namely the A-Not A method. Specifically, the goal of this experiment was to compare a ranking protocol with two A-Not A protocols. In all cases, there was more than one 'Not A' product. For one protocol, only 'A' was tasted beforehand for familiarization. For a second protocol, 'A' and all 'Not-A' products were given beforehand for familiarization. For a third protocol, the products were simply ranked.

2. Materials and methods

2.1. Judges

Seven experienced female panelists (age range, 45–61 y) were tested. Their experience of participating on sensory panels for testing margarines ranged 5–12 y. All were familiar with the A-Not A and ranking methods.

2.2. Stimuli

Six commercial margarines were obtained from the local supermarkets in Vlaardingen, Holland. These were: (A)

Halvarine (Gouda's Glorie, Zeewold, NL), (B) Havarine (Perfekt, Beesd, NL), (C) Bona, (Unilever Netherlands, Rotterdam, NL), (D) Volle Pond (Gouda's Glorie, Zeewold, NL), (E) Harvarine (C.I.V. Superunte B.A., Beesd, NL) (F) Sunflower (Gouda's Glorie, Zeewold, NL). For the purposes of this article, these products will be referred to by their corresponding letters 'A' to 'F'. All products were presented in 50 ml white plastic lidded cups under red light to minimize any color and reflectance differences. To sample the product, judges removed the lid and used separate plastic teaspoons for each tasting. Products were tasted and swallowed. Products were served chilled (5 °C) having been stored in a fridge until 5 min before serving. Between tastings, judges rinsed ad-lib with room temperature de-mineralized water (23–24 °C). Before beginning each of the three protocols, judges were allowed to eat Barber crackers (the horizon Biscuit Company Ltd., England) if desired; after this all judges then rinsed at least five times.

2.3. Procedure

Judges performed difference tests between the margarines using three separate protocols.

For the first protocol: 'ranking', product 'A' was presented to the judges as the standard. Judges were able to taste the standard as much as desired until they felt they had become familiar with its sensory characteristics (at least 4 teaspoonfuls). They were then given products 'A' to 'F' simultaneously and instructed to rank them in order of similarity to the standard. During testing, the standard and products 'A' to 'F' could be retasted as often as desired.

For the second protocol, a version of the A-Not A method was used. As before, product 'A' was presented to the judges as the standard. Again, judges were able to taste the standard as much as desired until they felt they had become familiar with its sensory characteristics (at least 4 teaspoonfuls). They were then given products 'A' to 'F' individually in random order, counterbalanced over sessions and required to report whether the products tasted the same or different from the standard. Responses were given in terms of six categories as follows: "same sure", "same not sure", "don't know, but guess it's the same", "don't know, but guess it's different", "different not sure", "different sure". During testing, the standard 'A' could be sampled as much as desired. For the purposes of this article, this protocol will be referred to as 'A-Not A: single'.

For the third protocol, a different version of the A-Not A method was used. As before, product 'A' was presented to the judges as the standard. However, this time products 'B' to 'F' were also presented simultaneously with 'A'. Judges were able to taste all these products as much as desired until they felt they had become familiar with the sensory differences between the standard 'A' and the products 'B' to 'F'. They were then given products 'A' to 'F' individually in random order and required to report whether they tasted the same or different from the standard. During testing, judges were not allowed to retaste the standard 'A' at will. For the purposes of this article, this protocol will be referred to as 'A-Not A: multiple'.

Judges performed all three protocols in a single session. They performed two sessions per day, lasting approximately 2 1/2 h, for a total of seven days (total 14 sessions). The order of presentation of the protocols was counterbalanced over sessions. There was a week interval between the first two days of testing. After a period of 10 months, testing the final 5 days was resumed at one week intervals. This schedule was determined by the limited availability of the trained taste panel for experimental work. However, examination of the data indicated that this unusual schedule did not adversely affect judges' performance.

2.4. Statistical analysis

R-Indices were computed in two ways: Firstly, for each product ('B' to 'F' as signals and 'A' as noise) and each protocol, R-Indices were computed individually for each judge (number of signals/noise = 14 per judge). Mean R-Indices, across judges but within protocols, were then calculated. For the second analysis, for each product ('B' to 'F' as signals and 'A' as noise) and each protocol, data for all judges were pooled onto a single response matrix and a single R-Index was computed. (number of signals/noise = 98 = 7 judges × 14 sessions).

3. Results and discussion

The computed R-Index values for products 'B' to 'F' for the three protocols are given in Table 1. As noted above, the two R-Index computations involving averaging judges' individual data and pooled data are also shown, as are means for all the products.

From the table, it can be seen that the highest R-Indices tended to be obtained with the ranking protocol. It may be hypothesized that this was because of the forced-choice nature of ranking and the results concur with previous research where ranking was seen to confirm Brown's (1974) prediction that higher R-Indices would be obtained with ranking than with a rating procedure (O'Mahony et al., 1980; Ishii et al., 1992).

The R-Indices for the 'A-Not A: multiple' protocol were higher than those for the 'A-Not A: single' protocol. It may be hypothesized that this was because the presentation of multiple standards gave the judges a better idea of the concept defined by the sensory characteristics of 'A'. Single presentation of 'A' would allow a concept to be formed, yet this concept could possibly be generalized so widely as to include some of the products from 'B' to 'F'. Yet, in the 'A-Not A: multiple' protocol, where the products 'B' to 'F' were presented beforehand along with 'A', this would allow the judges to form separate concepts for all these products. This would control the generalization of the concept for product 'A'. Thus, the boundaries of the concept for product 'A' would be better defined. This

Table 1
R-Index values (%) indicating differences between margarine products derived from ranking and two A-Not A methods, using two ways of combining data from individual judges

Method of combining judges' data	Products	Protocols			Grand total
		A-Not A: single	A-Not A: multiple	Ranking	
R-Indices calculated from pooled data	B	84.5	88.6	94.8	89.3
	C	84.3	88.2	94.0	88.8
	D	77.6	84.3	91.3	84.4
	E	76.6	80.5	90.5	82.5
	F	51.6	63.4	54.1	56.4
R-Indices calculated by averaging judges' R-Indices	B	84.9	92.2	96.1	91.1
	C	83.7	92.0	93.0	89.6
	D	75.4	90.0	90.5	85.3
	E	75.5	87.5	90.2	84.4
	F	51.1	68.3	54.8	58.1

would lead to fewer errors in the A-Not A test. For the single protocol, where such boundaries were not well defined beforehand, the concept of product 'A' would need to be established during testing. This would result in a higher error rate. This concurs with previous research that multiple standards, giving examples of stimuli both within and outside a sensory concept, provide a better definition of the concept than merely giving a single standard (Ishii & O'Mahony, 1991).

Next, it is interesting to compare the mean of the R-Indices computed from individual judges with pooled R-Indices where data from all judges are entered into a single matrix. In the latter case, where data from different judges are pooled on to the same matrix, judges would have different criteria. This would cause what is known as boundary variance. Boundary variance is a concept used in scaling. It refers to the fact that judges space their numbers differently when they are making numerical estimates using rating scales. Another way of describing this is to say that the boundaries between the numbers vary among judges. For example, judges will not place the boundary between numbers 6 and 7 at the same level of intensity. Thus, this boundary varies among judges, resulting in boundary variance. In the same way, with the A-Not A test, the boundaries between the categories 'sure' and 'not sure' and between 'not sure' and 'guessing' vary among judges. This added boundary variance has the effect of depressing sensory indices of difference. Another way of considering boundary variance is that one person's 'sure' is another person's 'not sure' and so entering both their data into the same matrix can result in more artificial ties and reversals. In the case where R-Indices were computed from individual judges, a judge would be expected to keep his own criteria fairly constant during an experimental session. Thus his individual R-Index values would not suffer from boundary variance and not be depressed. Thus the mean of such values would be expected to be higher than pooled R-Index.

For the ranking protocol, indices computed from pooled data and averaged from individual data did not show any systematic variation (*t*-test, $p = 0.99$). This is to be expected because for ranking the boundaries are fixed both within and between judges, consequently excluding boundary variance.

For the 'A-Not A: multiple' protocol, the effect of boundary variance in the pooled data is apparent. For all products, mean R-Indices were higher when calculated from individual judges (*t*-test, $p = 0.001$). The same effect would be expected for the 'A-Not A: single' protocol but it was not apparent (*t*-test, $p = 0.15$).

It would be difficult to argue that the lack of difference for the 'A-Not A: single' protocol was due to judges' all assuming the same boundaries (sure vs not sure vs guessing) as with ranking. Instead, it may be hypothesized that with only the single presentation of product 'A' beforehand, it was difficult for individual judges to establish stable boundaries. Thus, whether R-Indices were calculated from pooled data or by averaging judges' individual R-Indices, the boundaries would be unstable in both cases. Thus differences between the two computational methods would not appear and R-Indices would be depressed as seeing in Table 1.

It is worth returning to the hypothesis that the boundaries of the concept of product 'A' was better defined by prior presentation of all the products in the 'A-Not A: multiple' protocol. Such an argument implies a beta-criterion. However, if only product 'A' was presented beforehand, the judge may not be able to establish the boundaries of the concept (beta-criterion). He might be forced to use tau-criterion instead. If the test did not have sufficient replications, giving examples of 'A' and 'Not A', he would not gain enough conceptual information to establish beta-criterion. In this case, he would need to use a tau-criterion throughout testing. Thus, considering the limitation on number of replicates in sensory evaluation, the logical possibility exists that differences in the A-Not A protocols have the potential to induce different cognitive strategies. Furthermore, the A-Not A test has a commonality with the triangle, duo-trio, and same-different tests in that the attribute change is not specified; this later tests involve tau criteria. Only when the attribute is specified (2-AFC, 3-AFC), is the beta criteria involved.

Should the use of tau-criteria be the case that, an explanation for the difference between the 'A-Not A: multiple' and the 'A-Not A: single' protocols might be that for individual judges, tau-criteria are not as stable as beta-criteria. Therefore, the lack of stability of tau-criteria for individual judges in the 'A-Not A: single' protocol, would produce as much boundary variance as when data were pooled over judges.

Yet, more information is needed concerning the decision rules or cognitive strategies involved in the A-Not A test. It is not known at this point, whether slight differences in the instructions or procedure might elicit different cognitive strategies or only affect the perceptual learning process for establishing beta-criteria. It is also not known whether differences among judges in terms of their experience (prior familiarity) might not do the same. The latter is currently under investigation.

For products 'B' to 'E', R-Indices were higher for ranking than for the 'A-Not A: multiple' protocol than for the 'A-Not A: single' protocol. This was not the case for the product 'F' where R-Index values were close to 50% (chance level) except in the 'A-Not A: multiple' protocol. Obviously the difference between the product 'F' and 'A' was much smaller than other differences. It may be hypothesized that the reason that it was discriminated better by the 'A-Not A: multiple' protocol was that the 'familiarization' (prior presentation of 'A' and 'Not A' products) came closer to a warm-up procedure and thus elicited greater judge discriminability for such small difference (Dacremont, Sauvageot, & Duyen, 2000; O'Mahony, Thieme, & Goldstein, 1988; Thieme & O'Mahony, 1990).

The authors are aware that the ranking or categorization for similarity as in the A-Not A method does not use a univariate dimension; differences between products can be due to different attributes. As a tool, in sensory evaluation, such methods should be used with caution because rankings or categorization might depend on conceptual differences as well as sensory differences. For the present experiment, this was not an issue because comparisons were made using the same judges with the same idiosyncratic conceptualizations.

4. Conclusions

It was apparent from the data that the ranking elicited higher R-Index values than the A-Not A methods. It can thus be seen as more sensitive and useful replacement for the A-Not A methods, provided that a sensory panel is able to repeatedly retaste the products involved. Regarding the A-Not A methods, 'familiarization' (the prior presentation of all the test products) given to the judges before beginning the test seems to be important to stabilize the cognitive decision criteria (beta-criterion). It was hypothesized that when judges were not experienced with the test products enough to develop the concepts of all the test products, tau-criterion can also be used in A-Not A methods. For the R-Index, knowledge of the cognitive strategy is not necessary for its computation. However, differences in cognitive strategy can affect the level of performance and thus the R-Index. For example, an R-Index using a beta-criterion will be slightly higher than an R-Index using a tau-criterion (Noreen, 1981; Hautus & Irwin, 1995; Irwin & Francis, 1995).

Acknowledgements

The authors would like to thank Graham Cleaver for his advice on the manuscript and Mijntja Kersemaekers for her assistance with the experiment.

References

ASTM (1968). *Manual on sensory testing methods*. Philadelphia: American Society for testing and materials.

Brown, J. (1974). Recognition assessed by rating and ranking. *British Journal of Psychology, 65*, 13–22.

Casale, M., Ashby, F. G., & Standring, R. (2005). The role of perceptual learning in categorization. *Journal of Cognitive Neuroscience, Suppl. S.*, 123.

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.

Delwiche, J., & O'Mahony, M. (1996). Changes in secreted salivary sodium are sufficient to alter salt taste sensitivity: Use of signal detection measures with continuous monitoring of the oral environment. *Physiology and Behavior, 59*, 605–611.

Francis, M. A., & Irwin, R. J. (1995). Decision strategies and visual-field asymmetries in same-different judgments of word meaning. *Memory and Cognition, 23*, 301–312.

Green, D. M., & Swets, J. A. (1966). *Signal detecting theory and psychophysics*. New York: John Wiley & Sons.

Hautus, M. J., & Irwin, R. J. (1995). Two models for estimating the discriminability of foods and beverages. *Journal of Sensory Studies, 10*, 203–215.

Irwin, R. J., & Francis, M. A. (1995). Perception of simple and complex visual stimuli: Decision strategies and hemispheric differences in same-different judgments. *Perception, 24*, 787–809.

Ishii, R., & O'Mahony, M. (1991). The use of multiple standards to define sensory characteristics for descriptive analysis: Aspects of concept formation. *Journal of Food Science, 56*, 838–842.

Ishii, R., Vié, A., & O'Mahony, M. (1992). Sensory difference testing: Ranking R-indices are greater than rating R-indices. *Journal of Sensory Studies, 1*, 57–61.

Lawless, H. T., & Heymann, H. (1996). *Sensory evaluation of food: Principles and practices*. New York: Chapman & Hall.

Lee, H.-S., van Hout, D., Hutaus, M., & O'Mahony, M. (forthcoming). Can the same-different test use a beta criterion as well as a tau criterion? Food Quality and Preference.

Lee, H.-S., & O'Mahony, M. (2004). Sensory difference testing: Thurstonian models. *Food Science and Biotechnology, 13*, 841–847.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (second ed.). Newyork: Cambridge University Press.

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

Noreen, D. L. (1981). Optimal decision rules for some common psychophysical paradigms. In S. Grossberg (Ed.), *Mathematical psychology and psychophysiology*. *Proceedings of the symposium in applied mathematics of the american mathematical society and the* *society for industrial applied mathematics* (vol. 13, pp. 237–279). Providence, RI: American Mathematical Society.

O'Mahony, M. (1992). Understanding discrimination tests: A user-friendly treatment of response bias, rating and ranking R-index tests and their relationship to signal detection. *Journal of Sensory Studies, 7*, 1–47.

O'Mahony, M., Garske, S., & Klapman, K. (1980). Rating and ranking procedures for short-cut signal detection multiple difference tests. *Journal of Food Science, 45*, 392–393.

O'Mahony, M., Masuoka, S., & Ishii, R. (1994). A theoretical note on difference tests: Models, paradoxes and cognitive strategies. *Journal of Sensory Studies, 9*, 247–272.

O'Mahony, M., & Rousseau, B. (2002). Discrimination testing: A few ideas, old and new. *Food Quality and Preference, 14*, 157–164.

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.

Peryam, D. R. (1958). Sensory difference tests. *Food Technology, 12*(May), 231–236.

Pfaffmann, C., Schlosberg, H., & Cornsweet, J. (1954). Variables affecting difference tests. In D. R. Peryam, F. J. Pilgrim, & M. S. Peterson (Eds.), *Food Acceptance Testing Methodology* (pp. 4–20). Chicago: Quartermaster Food and Container Institute.

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.