ELSEVIER

0950-3293(95)00022-4

Food Quality and Preference **6** (1995) 227–238
Elsevier Science Limited
Printed in Great Britain
0950-3293/95 $9.50+.00

WHO TOLD YOU THE TRIANGLE TEST WAS SIMPLE?

Michael O'Mahony

Department of Food Science and Technology, University of California, Davis, CA 95616, U.S.A.

(*Received* 7 *Novemberr* 1994; *accepted* 22 *March* 1995)

ABSTRACT

The effects of response bias, cognitive strategy change and the sequence of tasting are discussed in relation to the triangle test. Theoretical approaches like Thurstonian modelling and Sequential Sensitivity Analysis are reviewed. It is concluded that the triangle test is prone to many pitfalls; even a slight change in the instructions can bring about a radical change in performance.

INTRODUCTION

Sensory difference tests are generally thought to be conceptually simple, yet there is a considerable amount of theory associated with them. This paper will examine some of the theory, models and pitfalls associated with difference testing using the triangle test as an illustration.

The triangle test (Peryam, 1958; Peryam & Swartz, 1950) is one of a set of triadic tests. Three food stimuli are presented to the judge, of which two are the same and one is slightly different. The judge is required to indicate the odd or different sample. In some versions of the test, the judge is asked to indicate the two samples which are the same (Helm & Trolle, 1946). Whether there is any difference in performance from these variations remains uninvestigated; logically they are the same, yet psychologically they could be different. Another triadic test is the 3-alterative forced choice or 3-AFC test (Green & Swets, 1966). This is just like a triangle test except the instructions specify the nature of the difference. For example, triangle instructions might be: "Here are three cakes, one is different from the other two; indicate the different one." Yet, 3-AFC instructions might be: "Here are three cakes, one has a coconut flavour while the other two have not; indicate the cake with the coconut flavour." A judge will tend to perform better on a 3-AFC than a triangle test, yet the reason why is not obvious. This will be discussed later, but first a basic problem with difference testing must be considered: response bias.

RESPONSE BIAS

Sensory difference tests are designed for the detection and measurement of fine sensory differences between foods or other stimuli. The differences are so small that the foods can easily be confused; special test procedures are required to determine whether they can be perceived as different or not. When judgements are being made of such small differences, factors come into play that we would not notice in ordinary life situations; one such factor is response bias.

Response bias can be illustrated by considering Fig. 1. On the left is represented an unsweet biscuit 'N'. As we travel towards the right of the figure, the biscuit becomes sweeter. At first the increase in sweetness is so small that it is difficult to determine whether it has occurred or not; this could be called *the region of uncertainty*. As the sweetness increases, it becomes more easily perceptible until we arrive at the sweet biscuit 'S', which is sweet enough to be always distinguished from 'N'. Difference tests would not be needed to determine whether 'S' and 'N' were different; the stimuli are not confusable.

Now consider the biscuit 'X'. It is not the same as the unsweet biscuit 'N'; it has a small amount of added sweetener. Yet, it is difficult to tell whether 'X' actually tastes sweeter than 'N' or not; it is in the region of uncertainty. Maybe it tastes sweeter, maybe it doesn't. Because of this uncertainty, difference tests are needed to determine whether 'N' and 'X' can be distinguished by sweetness. The two stimuli are confusable.

If a judge were asked whether 'S' was sweeter than 'N', he could respond easily. If he were asked whether 'X' was sweeter than 'N', he would be uncertain. In fact, the question implies a second question: How much sweeter than 'N' must 'X' be to be regarded as sweeter? The judge's response would depend on where 'unsweetness' finished and where 'sweetness' started; it would depend on where he 'drew the line' between 'unsweet' and 'sweet' (O'Mahony, 1989, 1992). If he were to 'draw the line' at 'D', sweetness would start at 'D' and so biscuits to the right of 'D' would be called 'sweet'. 'X' would be judged as sweeter than 'N'. If he were to draw the line at 'A', the biscuit 'X' would be

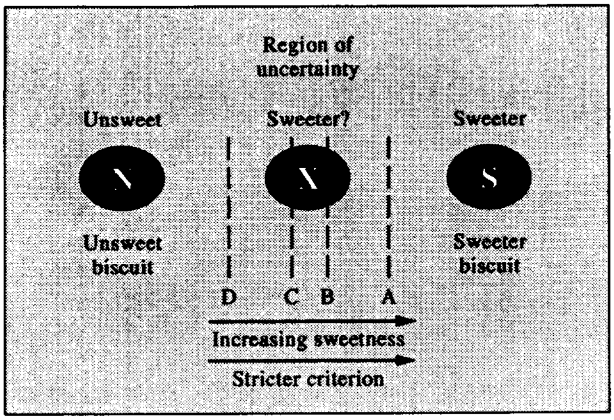

FIG. 1. Diagram representing response bias using an example of sweet and unsweet biscuits.

deemed to be 'unsweet' and so not different from 'N'. Where the judge 'draws the line' is his criterion of where sweetness begins. A biscuit must have this level of sweetness to be called sweeter than 'N'. Thus, a judge who is biased towards having his criterion at 'A' has a strict criterion, he is reluctant to call a biscuit sweeter and requires a high level of sweetness before he is prepared to commit himself to reporting it as such. If the judge is biased in the opposite direction and places his criterion at 'D', he would now have a more lax criterion. He would not need to have such a high level of sweetness before he committed himself to responding that 'X' was sweeter than 'N'. This tendency for a judge to have his criterion placed over to the right (strict criterion) or over to the left (lax criterion) is called *response bias*. Such response bias is a matter of how cautious a judge feels about committing himself to responding that 'X' has sweetness. It is a cognitive factor; it is independent of the sensitivity of the judge to sweetness.

In terms of Figure 1, if a judge were more sensitive to the sweetness of 'X', then it might appear easily distinguishable from 'N' and be placed over to the right with 'S'. The sweeter the biscuit appears, the more different it is from 'N', the further over to the right it will be in Figure 1. The argument can be generalised. The dimension considered in Figure 1 is the univariate dimension of sweetness. Yet, the same argument can be applied to a multivariate dimension with multivariate criteria to act as decision points for the judge's response.

The placement of the criterion is a cognitive factor depending on how cautious the judge feels. It can change arbitrarily, and this will change the judge's response. Yet, such changes are independent of the judge's sensitivity to the sweetness of the biscuit. It is important for any difference test protocol to adopt strategies that avoid this susceptibility to the effects of uncontrolled criterion shifts.

DEFEATING RESPONSE BIAS BY USING FORCED CHOICE PROCEDURES TO STABILISE THE CRITERION

There is a simple strategy for eliminating the effects of criterion shift, and it can be understood by referring to Figure 1. The strategy is to arrange things so that the criterion always falls between 'N' and 'X'. Then, if 'X' were detected as sweeter than 'N' it would always be reported as such. The criterion should not be allowed to shift to the right of 'X' or else 'X' would never be reported as sweeter. The challenge is to devise a testing protocol that stabilises the position of the criterion in the region between 'X' and 'N'.

In fact, it is remarkably easy to do this. It is achieved by all the common forced-choice procedures, like the directional paired comparison, duo–trio, or triangle test (Helm & Trolle, 1946; Peryam, 1958; Peryam & Swartz, 1950).

As an example, consider a directional paired comparison (sometimes called a 2-alternative forced-choice or 2-AFC; Green & Swets, 1966) where the judge is given the two biscuits and is told to select the sweeter of the two. The judge does not have to decide whether a sensation that might be sweetness in one biscuit, is sweet enough to be called "sweet"; essentially, he is told that it is. The judge knows that one biscuit will be more sweet, the other will be less sweet. The biscuit that has more of a sensation that may be construed as sweet must, therefore, be judged as sweeter, whether the judge feels cautious or reckless. In effect, the criterion sensation for sweetness is forced to fall somewhere between the sensations elicited by the two biscuits, 'N' and 'X'. The test instructions stop it straying; straying could result in both samples to be reported as sweet (or not sweet); it is essentially stabilised in a position between 'X' and 'N'.

It is worth noting that a slight change in the instructions could wreck the stabilisation of the criterion. If the judge were only told that one of the biscuits might be sweeter than the other, the judge could then possibly place his criterion to the right of 'X'. The sensation elicited by 'X' would then be regarded as not sufficiently different from the sensation elicited by 'N' to be regarded as 'different', even though the two sensations could be distinguished. Lack of control of the position of the criterion would have caused a detected difference not to be reported.

Another variation in the instructions might be to ask whether 'N' and 'X' were the same or different. Again, such instructions do not stabilise the criterion in the correct place. Unlike the instructions for the directional paired comparison, they do not imply that 'N' and 'X' must be on separate sides of the criterion. The answer would clearly depend on whether the criterion were shifted to the left or to the right of 'X'. Such instructions are used in the same–different test, which can thus be seen to be prone to response bias. The same–different test should not be used for difference testing in this form.

Similarly, for the triangle test, the judge does not have to decide whether a given sample is sufficiently different from the other two, to be called 'different'. The criterion is adjusted so that one sample is called 'different', the other two 'same'. Should there be two 'N' samples and one 'X', the criterion would be forced between 'N' and 'X', so that if the judge could detect a difference, the position of his criterion would allow him to report it. If triangle instructions were altered and the judge only told that one of the biscuits *might* be different, the judge would no longer be forced to place his criterion between 'X' and 'N' and the test would again be prone to response bias.

It can be seen that an understanding of response bias leads to a realisation of how important precise instructions are for forced-choice procedures. It is essential for any forced-choice procedure that the judge should know the number of stimuli lying on different sides of the criterion. Without this knowledge, the criterion cannot be stabilised in the appropriate position. In the paired comparison, the judge is told that one stimulus falls on each side of the criterion. In the triangle or 3-AFC, the judge is told that one stimulus falls on one side of the criterion while the other two fall on the other side. To be more precise, the judge, in essence, is told in the 3-AFC that the criterion must fall between the strongest and the two weaker stimuli (or vice versa). In the triangle, the judge, in essence, is told that it may fall between the strongest and the two weaker stimuli or between the weakest and the two stronger stimuli, but that it can only fall in one of these two places. For both the 3-AFC and triangle tests, this is sufficient stabilisation to overcome response bias. This point will become clearer after a consideration of Thurstonian modelling later in this paper.

The forced-choice procedure which stabilises the criterion is only one approach to solving the problem of response bias. Another approach is to use multiple criteria and from the variation in response over these criteria compute an index of sensitivity to the difference; this is the signal detection approach and it has been reviewed in detail elsewhere (Green & Swets, 1966; O'Mahony, 1992).

THE PARADOX OF DISCRIMINATORY NONDISCRIMINATORS AND THURSTONIAN MODELLING

It can be seen that the instructions for a triangle test must be put in such a way as to stabilise the criterion. If the judge were told that one of the stimuli *may* be different from the other two, rather than it *will* be, response bias would no longer be controlled and the judge's performance would deteriorate. There are other changes in the instructions for a triangle test that could occur and these can improve performance. If the nature of the difference were specified (and now the test would be called a 3-AFC), performance improves. Byer and Abrams (1953) first noticed this, and the 'paradox' was later called the *paradox of discriminatory nondiscriminators* by Gridgeman (1970). Various studies have confirmed the effect (Frijters, 1981; Geelhoed *et al.*, 1994; MacRae & Geelhoed, 1992; Raffensberger & Pilgrim, 1956; Stillman, 1993; Tedja *et al.*, 1994). The effect is best understood by considering Thurstonian modelling.

Ura (1960) first applied Thurstonian (1927*a*, 1927*b*) ideas to difference tests, and although the explanation of the paradox was implicit in his work and that of later authors (David & Trivedi, 1962), it took the pioneering work of Frijters (1979, 1981; Frijters *et al.*, 1982) to draw together the theory and explain the paradox. Frijters was the first to use Thurstonian arguments to point out that the 'paradox' was not really a paradox, but simply a difference in performance elicited by the use of different cognitive strategies for the triangle and 3-AFC methods. Since then, tables of d' have been published, using decision rules based on these strategies, for the proportion of tests correct for triangle and 3-AFC tests (Ennis, 1993; Frijters, 1982; Frijters *et al.*, 1980).

The basic idea behind Thurstonian modelling is that each time a product is tasted, it will vary in its flavour intensity. This can be a result of physiological effects like sensory adaptation, or it can even be due to lack of homogeneity in the samples of the products themselves. The precise reasons are generally not pursued by the modellers, although an exception is provided by Ennis and Mullen (1992*a*) in which stimulus noise is separated from neural noise. Yet usually, for most modelling studies, it is sufficient that the flavour intensity of a stimulus merely varies. Sometimes the stimulus will taste stronger, sometimes it will taste weaker. There will be an average intensity which will occur most commonly; stronger and stronger intensities will occur less and less commonly, as will weaker and weaker intensities. Such variation in flavour intensity can be represented by a continuous frequency distribution along a flavour intensity axis (see Fig. 2), whereby the height of the distribution represents how commonly each intensity will occur. The height is the greatest at the most commonly occurring mean intensity but becomes lower at the less frequent high and low intensities. The momentary intensity upon tasting will be some value along the axis; how commonly that value occurs will be represented by the distribution. Thurstonian modellers are interested in the shape of the distribution; they are interested in trying various distributions and seeing which best fit the data. The choice of the normal distribution is most

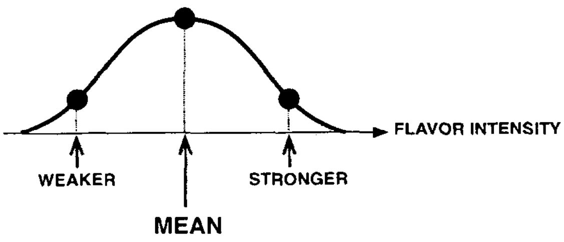

FIG. 2. Frequency distribution along a flavour intensity axis representing variation in flavour of a stimulus on repeated tasting. The axis can be univariate, multivariate or involve multiple sensations.

common, although other, non-normal distributions have been used (Frijters *et al.*, 1980). The intensity axis represented in Figure 1 is univariate; it could just as easily be multivariate, to reflect the multivariate nature of sensory stimulation given by a food. It could even be multiquality, with different areas of the axis giving different sensations; this can occur for such stimuli as low concentration NaCl solutions. The exact nature of the axis does not affect the arguments given here.

Consider Figure 3. In part (a) are represented the flavour intensity distributions for two food samples: 'N' and 'S'. Food 'S' has a stronger mean flavour intensity than food 'N' and, accordingly, is further up the flavour intensity axis (towards the right). The distribution for the stronger food, 'S', is so far up the intensity axis that it does not overlap with the distribution for the less intense food, 'N'; the two are quite separate. This means that even at its lowest intensity, food 'S' will always have a stronger flavour than food 'N'. There is no overlap; the foods are easily distinguishable.

Consider part (b). The mean flavour intensity for 'S' is still greater than that of 'N', but the distributions are now closer; they overlap. This means that the intensity of 'S' at its lower values is actually less than that of 'N' at its higher values. This means that there will be occasions when 'N' can actually appear to taste stronger than 'S'. The two foods are similar enough to be confusable.

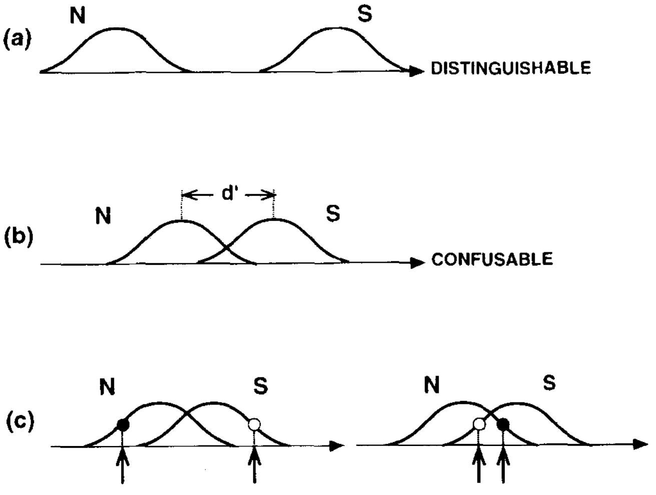

FIG. 3. Frequency distributions representing distinguishable and confusable stimuli, with a Thurstonian modelling approach illustrating correct and incorrect paired comparisons for the confusable stimuli.

Obviously, the more the two distributions overlap, the more the two foods will be confused. The less the two distributions overlap (the greater the distance between the means), the more distinguishable the two foods will be. A simple measure of overlap is d' (some authors use δ because δ is a population parameter). This is the distance between the two means, measured in units of standard deviation. Two distributions whose means are two standard deviation apart ($d' = 2$) overlap less and will thus be more easy to discriminate than two whose means are only one standard deviation apart ($d' = 1$). This is true, however big or small the standard deviations may be. The greater the standard deviation of the distributions (the variation in flavour intensity), the further apart the means must be for the foods to be distinguished. To simplify matters, the assumption is generally made that the two distributions have the same standard deviation, so that there is no confusion about which distribution is to provide the standard deviation units for d'. Should they have different standard deviations, another symbol (Δm) is used along with a parameter specifying the relative size of the standard deviations (Green & Swets, 1966).

Considering Figure 3 part (c), one can now see the results of such confusability between the food stimuli, as represented by the overlapping distributions. Consider the case on the left. At the instant 'S' is tasted by the judge, the stimulus happens to evoke a high intensity value; it is at a high point along the intensity axis. The instant 'N' is tasted, this second stimulus happens to evoke a low intensity value; it is at a lower point along the intensity axis. Accordingly, 'S' will taste stronger to the judge than 'N'. If asked to identify the stronger of the two foods, the judge would correctly choose 'S'. Now consider the case on the right. At the instant of tasting, 'S' is at a lower intensity, while 'N' is at a higher intensity. This time, 'N' actually feels stronger than 'S'. If asked to identify the stronger flavoured food, the judge will choose 'N'. Yet, 'S' is meant to be the stronger food and the judge will be treated as having made an error.

The test described above is, of course, the paired comparison test (Peryam, 1958), sometimes called the 2-AFC or 2-alternative forced choice test (Green & Swets, 1966). By choosing appropriate shaped distributions (usually normal) and a given d', Thurstonian modellers can select flavour intensity values for the two foods according to the likelihoods dictated by the two distributions. Each time two values are selected, the modeller can see whether the test would have been performed correctly. Repeating this process thousands of times, the proportion of paired comparisons that would be performed correctly for a given d', can be determined;

this process is called Monte Carlo simulation. Tables can be generated in this way giving values of d' for given proportions of correct paired comparisons. Using this type of modelling or even more analytic approaches, tables have been published using this approach for a range of sensory tests (Elliott, 1964; Ennis & Mullen, 1986*b*; Frijters, 1980; Frijters *et al.*, 1980; Hacker & Ratcliffe, 1979; Ura, 1960).

The question posed in the paired comparison test refers to variation along a unidimensional axis (pick the sweeter, crunchier, fruitier etc.). However, foods often vary on more than one attribute. A question specifying a difference on just one of these attributes would then be inappropriate if the aim was to allow the judge to choose from all the attributes, when making his discrimination. The problem is intensified should the experimenter not know beforehand which attributes might be varying. The solution involves a technique called 'warm-up'. For this, the two stimuli to be discriminated are presented to the judge immediately before testing. The judge tastes the two stimuli alternately. At first, a difference might not be apparent, but after several tastings of each stimulus, the difference generally begins to be perceived. This is probably a function of the attention mechanism managing to sort out the signals specifying the difference from the background noise. It appears that judges can 'warm up' on many foods, and such a procedure before testing increases discrimination performance (O'Mahony *et al.*, 1988). After warm-up, a judge can then be asked to think of a descriptive term to describe the difference that was perceived during warm-up. The descriptor might describe a set of differences perceived analytically as subjectively separate or perceived as a blending of attributes synthesised into a singular appearing attribute; it does not affect the technique. The descriptor chosen by the judge can then be used to pose the paired-comparison question. Such a procedure is called the 'warmed-up paired comparison' (Thieme & O'Mahony, 1990) and provides a way of performing a paired comparison test when a descriptive term to describe the difference is not readily available, or the nature of the difference is not known.

These arguments can now be applied to triadic tests: the triangle and the 3-AFC.

Consider Figure 4. Let us assume that in this case, the triad has two 'N' samples and one 'S'. In part (a), at the instants the two 'N' samples are tasted, 'N' is at its lower intensities. When 'S' is tasted, it is at a higher intensity. If asked to indicate the strong food (3-AFC instructions), the judge will correctly identify 'S'. If asked to indicate the odd or different food, the judge will also correctly, identify 'S'. In both cases, the judge would be correct.

Consider part (b). When the two 'N' samples are tasted, one is at a high intensity while the other is low. The 'S' sample, when tasted, is at a low intensity, lower even than the 'high N'. When asked to indicate the odd sample, the judge will choose the 'low N' because it is furthest away along the axis from the other two. When asked to indicate the strongest sample, the judge will choose the 'high N', being the furthest up the intensity axis (furthest to the right). In both cases (3-AFC, triangle), the judge will be treated as having made an error. It should be noted that the 'N' chosen in the triangle test is different from the one chosen in the 3-AFC.

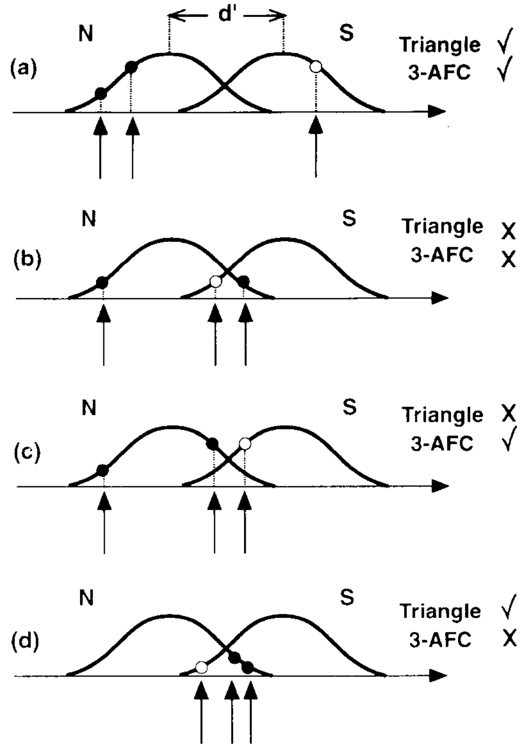

FIG. 4. Thurstonian treatment of correct and incorrect triangle and 3-AFC tests, indicating the paradox of discriminatory nondiscriminators.

Considering part (c), the judge, when asked to indicate the odd sample, will choose 'N' at its low intensity (the one furthest to the left), because it is a greater distance away from the other two along the axis. Thus, with triangle instructions, the judge will be deemed to have made an error. Yet, with 3-AFC instructions, the judge will choose the strongest sample (the one furthest to the right): 'S';. Here, the judge will be scored as correct. So, there are occasions when the 3-AFC will be performed correctly and the triangle incorrectly.

Considering part (d), the judge, when asked to indicate the odd sample, will choose 'S' at its low intensity because it is a greater distance away from the other two along the axis. Thus, with the triangle instructions, the judge will be scored as correct. When given the 3-AFC instructions, the judge will choose the 'N' with the highest intensity and will be scored as having made an error. Here, the triangle will be performed correctly and the 3-AFC incorrectly.

So, there are occasions when the triangle and 3-AFC tests are both performed correctly (see Fig. 4a) and when they are both performed incorrectly (see Fig. 4b). On some occasions, the 3-AFC will be performed correctly and the triangle incorrectly (see Fig. 4c), on others the reverse will be the case (see Fig. 4d). However, the case of more 3-AFCs being performed correctly (Fig. 4c) will occur more frequently than the case where more triangles are performed correctly (Fig. 4d). A look at Figure 4 will indicate that the case where both 'N' samples are stronger than 'S' (Fig. 4d) is likely to be quite rare compared with a case when one of the 'N' stimuli happens to be closer to the 'S' stimulus (Fig. 4c). Also, the heights on the frequency distributions in Figure 4d are smaller than on Figure 4c, indicating that it will occur less frequently. This means that overall, a judge will perform a higher proportion of 3-AFC tests correctly than triangle tests. The judge's sensitivity (d') will not have changed; there will still be the same degree of overlap between the two distributions. What has changed is the decision rule or the cognitive strategy used to find the target stimulus.

In the triangle case, the judge is required to compare the distances along the flavour intensity axis between the intensity values for the three food samples [see Fig. 5, part (a)]. He then selects the sample that is most separate in distance from the other two. This is called the 'comparison of distances' strategy (O'Mahony *et al.*, 1994). To be more specific, the strategy involves a comparison of the distances between each pair of stimuli. Each stimulus has two distance values, each one signifying the distance between it and one of the other two stimuli. The stimulus whose two distance values adds up to the highest total is the stimulus that is the furthest away from the other two.

In the 3-AFC test, the judge is merely required to select the stimulus with the highest intensity value, the one that is furthest to the right along the axis [see Fig. 5, part (b)]. The strategy can be seen as moving down the axis from the right and selecting the first (highest intensity) stimulus encountered. It is rather like skimming off the most intense stimulus, like skimming the cream off the top of the milk. Accordingly, it has been called a 'skimming' strategy (O'Mahony *et al.*, 1994). Of course, there may be one weak and two strong stimuli, in which case the judge will select the lowest intensity value.

This difference in strategy or decision rule accounts for the difference in performance; the extent of the difference can be seen in Table 1. Here, the percentage of correct responses for the triangle and the 3-AFC tests have been computed for various d' values, using Thurstonian modelling (Ennis, 1993*a*). It can be seen that the percentages are higher for the 3-AFC test. For $d' =$ 1·5, a judge is most likely to get 76·58% of 3-AFC tests correct but only 50·56% for the triangle test. If this judge were to perform 10 tests, he would be most likely to get 5 triangle tests correct. Yet, he would be most likely to get between 7 and 8 3-AFC tests correct. At p = 0·05 using binomial statistics (Roessler *et al.*, 1978) he would be judged as discriminating significantly on the 3-AFC test but not on the triangle test. Yet, the judge would be equally sensitive to the differences between the foods in both cases; d' would not have changed. The trouble is that, because the search strategy used by the judge in each case is different, the 3-AFC and triangle tests are not equivalent. They may be statistically equivalent when the judge is guessing (both have a 1/3 chance of being guessed correctly), but they are not equivalent when subjects are not guessing. The judge is not doing the same thing in the 3-AFC as in the triangle test; the standard binomial tables used for difference tests cannot take account of this and are only set up to see how well a guessing model (null hypothesis) can be rejected. The trouble is that the judge is not guessing; he is using one of the cognitive strategies illustrated in Figure 5. Furthermore, binomial tables used for difference testing have nothing to say about the extent of the differences found. Reluctantly, it can be seen that tables for difference tests based on binomial statistics have some shortcomings and that measures of degree of difference or sensitivity like d', or its nonparametric equivalent: the R-index (O'Mahony, 1992), would be preferable.

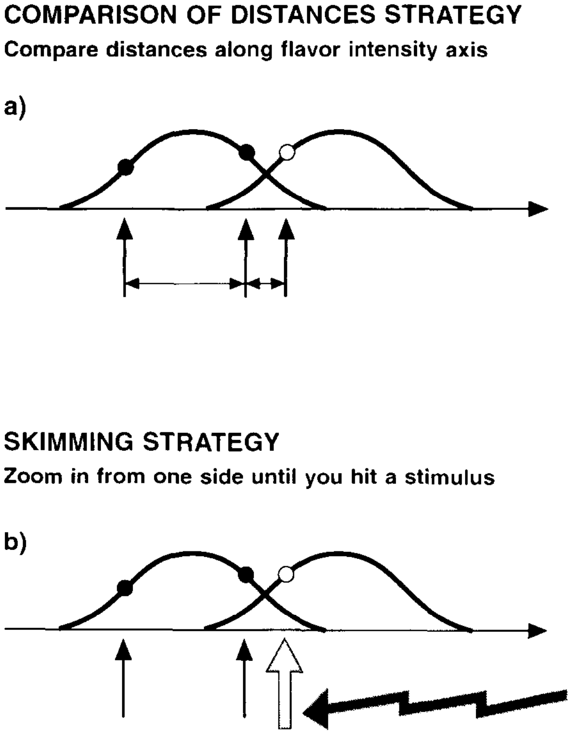

FIG. 5. Illustration of two cognitive strategies or decision rules used during difference tests: comparison of differences vs. skimming.

One common explanation of why judges perform better on 3-AFC than a triangle is that they know what they are looking for. Yet, this is not the reason for the superior performance. The fact that the judge knows the

TABLE 1. Proportion of Correct Responses for the Triangle and 3-AFC Methods for Given d' Values

d'	Triangle	3-AFC
0·0	33·33%	33·33%
0·5	35·58	48·26
1·0	41·80	63·37
1·5	50·56	76·58
2·0	60·48	86·58
2·5	69·93	93·14
3·0	78·14	96·88

nature of the difference allows him to use a 'skimming' strategy rather than a 'comparison of differences' strategy. The 'skimming' strategy elicits superior performance with triadic testing, but this is not necessarily true with other protocols. The point is illustrated by considering the tetrad test (Byer & Abrams 1953; Gridgeman, 1956; Lockhart, 1951). For this test, two sets of identical stimuli, two 'S' samples and two 'N' samples, are presented to the judge who has to distinguish between them by sorting them into two groups of two (two 'S' samples in one group, two 'N' samples in the other). The instructions to the judge can be given in two ways. Firstly, "Here are four drinks. They are, in fact, two sets of two identical drinks. Taste them and sort them into two groups of two. In each group the two drinks should be identical." This should induce the 'comparison of distances' strategy. The second instructions specify the nature of the difference and so induce the 'skimming' strategy: "Here are four drinks. Two have a citrus flavor; two have not. Taste them and identify the two citrus flavoured drinks." Unlike with triadic methods, the 'skimming' strategy does not induce superior performance over the 'comparison of distances' strategy. The explanation for this in terms of Thurstone models has been fully discussed elsewhere (O'Mahony *et al.*, 1994) and so will not be examined in detail here. Suffice it to say that specifying the nature of the difference between the stimuli (inducing the skimming strategy) does not always elicit superior performance; it merely induces a strategy that elicits, superior performance for some methods (e.g. triadic) but not for others (e.g. tetradic). So judges cannot be said to perform better in the 3-AFC simply because they have been told what to look for; it is because knowing what to look for elicits the more favourable 'skimming' strategy.

Thus, it can be seen that the cognitive strategy used to examine the input from the senses can alter performance, irrespective of sensitivity; Thurstonian models provide a good explanatory tool for this. Yet, there are other effects that actually alter d'; these are caused by carry-over effects in the mouth and are described by another model, Sequential Sensitivity Analysis or S.S.A.

CARRY-OVER EFFECTS AND SEQUENTIAL SENSITIVITY ANALYSIS

The Thurstonian modelling approach makes the assumption that the two distributions representing the two stimuli to be discriminated, are independent. This implies that the flavour of a stimulus is not affected by the flavour of the preceding stimulus. With suitable presentation procedures and interstimulus protocols (Halpern, 1986; O'Mahony, 1979) it is possible to achieve this. Yet, generally there will be carry-over effects from one stimulus to the next. So strong are these carry-over effects that an ordinal model based on them can be used to predict judges' ability to discriminate on a variety of tests. Because the model considers the sequence of tasting in any given test protocol, and how the after-effects of tasting one stimulus can affect or even obscure the taste of the following stimulus, the model is called Sequential Sensitivity Analysis (S.S.A.).

Consider two stimuli to be discriminated; one (S) is stronger in some or indeed many attributes, while the other stimulus (W) is weaker. In any sequence of tasting, there are four possible paired sequences that can occur: the strong stimulus following the weak (W–S), the weak following the weak (W–W), the weak following the strong (S–W) and the strong following the strong (S–S). Thus, a triadic test having the sequence W–W–S has the two pairs W–W and W–S, while the sequence S–S–W has completely different pairs: S–S and S–W. A strong stimulus tasted immediately after a preceding identical strong stimulus (S–S) will not taste as strong as after a weak preceding stimulus (W–S). This is because the first stimulus will adapt the sensory system and the resulting loss of sensitivity will make the second stimulus taste less strong. The stronger the first stimulus, the stronger the adaptation effect and the weaker will be the sensation from the following stimulus (O'Mahony, 1979, 1986; O'Mahony & Heintz, 1981; O'Mahony & Wingate, 1974). Actually, carry-over effects encompass more than just sensory adaptation, as will be discussed later. Yet, however complex the carry-over effects might be; they can easily be measured.

Consider a judge given a discrimination task to discriminate between food 'W' and food 'S' which is stronger in one or several attributes. The judge is given a random order of 'target' food samples and is requested to state whether they are 'S' or 'W'. The judge must have some knowledge of these foods to be able to know what signals to look for; this might have been acquired by 'warm up' (O'Mahony *et al.*, 1988). If, before tasting each target food sample, the judge also tastes a sample of 'S' or 'W' (also in random order), ignores its taste and then goes on to taste the target sample, the judge will then have been given a series of 'S' samples tasted after either 'S' or 'W' (S–S, W–S) and 'W' samples tasted after either 'S' or 'W' (S–W, W–W). Performance on how well he correctly names the target samples, 'S' and 'W', in each of these four sequences (W–S, S–W, W–W, S–S) is then measured. There are some complications to such a task, which, as stated above would be prone to response bias. Accordingly, the judge is made to respond using a signal detection rating task to circumvent response bias. All he has to do is state whether he is sure or not of his judgment. He can thus respond "W-sure", "W-unsure", "S-sure", "S-unsure". From these data, *R*-index measures of his performance (O'Mahony, 1992) can be computed to give a measure of the relative signal strengths of 'S' tasted after 'W' or 'S' and of 'W' tasted after 'W' or 'S'. The experimental techniques have been outlined in detail elsewhere (O'Mahony & Goldstein, 1986, 1987; O'Mahony & Odbert, 1985; Tedja *et al.*, 1994; Vié & O'Mahony, 1989) and so will not be detailed here. From such measurements, a definite order seems to emerge. Best identification is achieved for a strong stimulus following a weak one, while worst is for 'strong' following 'strong'. Second best is 'weak' following 'strong' and third is 'weak' following 'weak'. So the order can be listed as follows:

W–S	best
S–W	
W–W	
S–S	worst

Given this order, predictions can be made for triadic testing. The prediction is that a triad with the strong stimulus as odd will be easier to discriminate than a triad with the weak stimulus as odd; this can be understood from Figure 6. The figure shows at the top the three orders of presentation that can occur with a triad having the strong stimulus as odd. The next part shows the three orders that can occur when the weak stimulus is odd. The bottom part shows the relative ease of identification of the stimuli: a strong stimulus tasted after a weak stimulus (W–S), a 'weak' after a 'strong' (S–W), 'weak' after 'weak' (W–W) and lastly 'strong' after 'strong' (S–S). The number of each of the paired sequences occurring in the first three (strong-odd) and in the second three (weak-odd) can now be considered. Both have two (W–S) sequences, so the scores at the bottom are 2 and 2. Both have two (S–W) sequences, so again, the scores at the bottom are 2 and 2. The sequences with the strong stimulus as odd have two W–W sequences; the sequences with the weak stimulus as odd have none. In this case, score 2 for the strong-odd triads and zero for the weak-odd triads. The reverse can be seen to be true for the worst sequence (S–S). The weak-odd triads have two such sequences, the strong-odd have none. So, overall, looking at all the sequences, the triads with the strong stimulus as odd have a more distinguishable set of sequences than the triads with the weak stimulus as odd; their scores tend more towards the more distinguishable pairs. It should be easier to discriminate in a triadic test with the strong stimulus being odd than when the weak stimulus is odd.

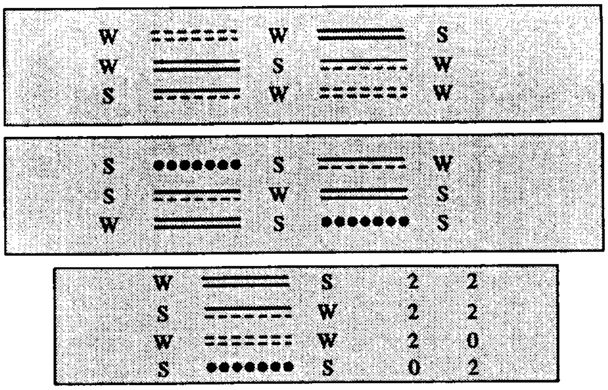

FIG. 6. Representation of the distinguishability of weak and strong stimuli in the three triads that have the strong stimulus as odd and the three triads that have the weak stimulus as odd.

The explanation given for Figure 6 is not complete. The first stimuli in the triads also followed something. It may have been a mouthrinse taken between triads or it may have been the final stimulus in the previous triad. When these possibilities are taken into account, the prediction for the triads remains the same.

The predictions regarding 'strong odd' triads and 'weak odd' triads have been confirmed in studies where the sequence has been measured as: W–S, S–W, W–W, S–S. For the same judges, performance on triads with the strong stimulus as odd has been superior. It has been confirmed for distinguishing low concentration NaCl (S) from distilled water (W) and a sweetened cherry flavoured beverage (S) from unsweetened beverage (W) (O'Mahony & Goldstein, 1986, 1987; O'Mahony & Odbert, 1985; Tedja *et al.*, 1994; Vié & O'Mahony, 1989). The same approach has also been applied to duo–trio tests and paired comparisons (O'Mahony & Odbert, 1985; Thieme & O'Mahony, 1990).

It is important to point out that this is an average effect. The sequence: W–S, S–W, W–W, S–S is the mean order obtained from a group of judges. Individual judges may vary, and also a given judge can change over time. With enough practice, a judge can become skillful enough to judge each sequence equally well, so that S–S can be identified, as well as W–S (O'Mahony & Goldstein, 1987).

The reason why the pairs are ordered as: W–S, S–W, W–W, S–S, has been investigated for NaCl and water stimuli and it is due to a number of effects: sensory adaptation, differential supra- and subadapting taste sensitivity, dilution of stimuli by saliva and response bias (O'Mahony & Goldstein, 1987). Such an order for foods is expected to be the result of equally complex interactions.

It is important not to generalise too far. The paired sequence order (W–S, S–W, W–W, S–S) has only been found in two systems: NaCl vs. water, sweetened vs. unsweetened cherry beverage. In some measurements with wine, judges were not discriminating, so all stimulus pairs were equivalent (O'Mahony & Goldstein, 1986). Current work in our laboratory is examining the order of sequence pairs for yoghurt tasting; in this case it may be that the major effect is merely a lack of discriminability in the S–S sequence which itself could account for the superior performance on 'strong-odd' triads.

However, the prediction of superior performance on a triad with the stronger stimulus as odd, has generally (though not always: Frijters, 1981) been confirmed (Filipello, 1956; Frijters *et al.*, 1982; Grim & Goldblith, 1965; Helm & Trolle, 1946; Hopkins, 1954; O'Mahony & Odbert, 1985; Stillman, 1993; Tedja *et al.*, 1994; Wasserman & Talley, 1969).

So S.S.A. is only just beginning to be developed as a model, but its simplicity is appealing. However, the effect that S.S.A. explains, that judges on average tend to discriminate triads better when the strong stimulus is odd, is well established. Triadic tests give varying d' values depending on the sequence of tasting.

TOWARDS A NEW MODEL

The traditional Thurstonian approach assumed no carry-over effects between stimuli. This is a sensible assumption when working with visual stimuli, but the assumption breaks down when it is applied to food and the chemical senses, where carry-over and adaptation effects can be strong enough to make a sensation vanish completely. So strong are the carry-over effects that a predictive model, S.S.A., can be based upon them.

One way of bringing these carry-over effects into the Thurstonian fold is to say the signal from a stimulus is conditional on the stimulus that preceded it. Thus, there would not be two distributions: one for the strong stimulus (S) and one for the weak (W). Instead, there would be four distributions: one for a strong stimulus preceded by a weak (W–S), one for a 'strong' preceded by a 'strong' (S–S), one for 'weak' preceded by 'strong' (S–W) and one for 'weak' preceded by 'weak' (W–W). Such a model is illustrated in Figure 7.

On the top row of Figure 7, the traditional Thurstonian picture of two distributions (for 'W' and for 'S') is given on the right. On the left, the new picture is represented, a distribution each for: W–W, S–W, S–S, W–S. This is simply saying that a stimulus preceded by a weak stimulus is not the same as when it is preceded by a strong stimulus; they should have separate distributions.

Consider the second row. The triad W–W–S is considered. Also, the stimulus that was tasted before the first stimulus in the triad, was another W, so the whole tasting sequence was W–W–W–S. An example of such a sequence would be a triad with two water stimuli followed by a salt stimulus, with a water rinse taken beforehand. In the traditional way, this situation would be dealt with by taking two stimuli from the W distribution and one from the S distribution. This is pictured on the right. With the new model, pictured on the left, there would be two stimuli from the W–W distribution and one from the W–S distribution. For both the old and the new approach, only two distributions would be necessary.

Now consider the bottom row, the triad S–W–W, with a W-sample taken beforehand. The traditional two distribution approach, given on the right, would be the same as in the first example. Yet, with the new approach, three distributions would be involved: W–S, S–W, W–W. In this case, the new model deviates from the old model. So the new model is more complex than

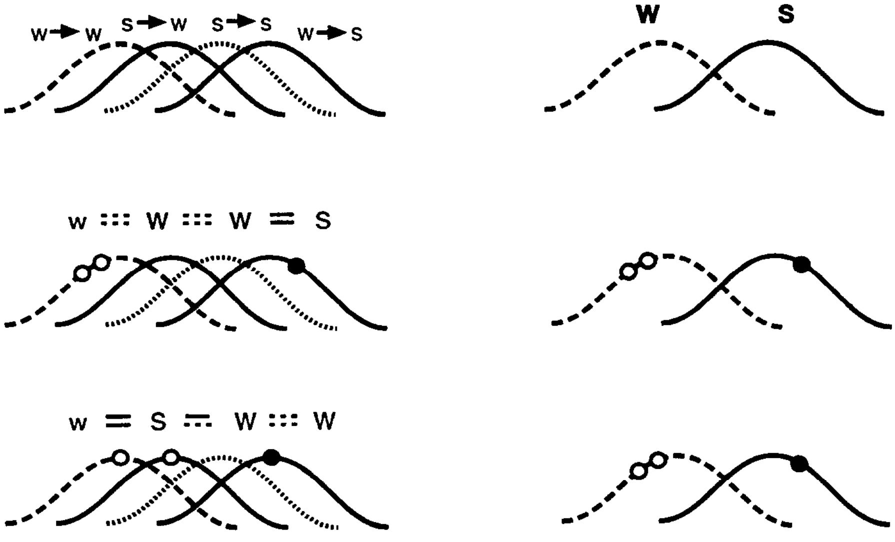

FIG. 7. Representation of two triads on a two distribution and a four distribution model.

the old. Either two or three of the four distributions may be used to represent a given triad, depending on the stimulus sequence in that triad. It also turns out that for triads where the weak sample (W) is the odd sample, the distributions required tend to be closer; they overlap more. This allows a greater chance of confusion, thus hindering performance. In this way, the new model predicts that triads with the weak stimulus as odd give inferior performance.

Thus, considering Thurstonian ideas, it would seem that the usual binomial analysis of number of tests correct is wanting. Obtaining measures of d' is better. Yet, now when we consider carry-over effects, even a measure of d' between two distributions is inadequate, because there should be four distributions. So how do we proceed?

A four distribution model was suggested by the experimental work of Tedja *et al.* (1994). A four distribution model has been developed (Ennis & O'Mahony, in press) to explain these data. It is complex and requires more than one d', but it was successful in explaining sequence effects, variation of cognitive strategies and even position bias. However, it is still possible to make good predictions using the components of the new model separately, namely S.S.A. and two distribution Thurstonian reasoning. Another approach which becomes obvious from Figure 8 is to take the 'S' distribution as bimodal (combine S–S and W–S), and also the 'W' distribution (combine S–W and W–W). With two bimodal distributions rather than two normal distributions, a nonparametric measure of discriminability like the *R*-index (O'Mahony, 1992) can be used.

CONCLUSIONS

It would seem that the triangle test is not as simple as it first appeared. There is position bias (Berg *et al.*, 1995; Frijters, 1977; Harries, 1956; Harrison & Elder, 1950; McBride & Laing, 1979; Tedja *et al.*, 1994), which not surprisingly, varies with the judges tested and can be countered by randomising or counterbalancing the order of presentation of the stimuli in the triad.

More serious is the effect of order of tasting which alters d', the ability of the judge to discriminate between the two stimuli. Discrimination is generally better in a triad with the stronger stimulus as the odd one. This allows a triadic test to be manipulated to increase or decrease its sensitivity. Yet, unless this is fully understood, the test may be manipulated inadvertently.

Then there is the importance of controlling response bias. As long as the instructions indicate the number of stimuli on each side of the criterion, response bias is controlled. Yet, if the instructions are varied slightly to: "one of the stimuli *may* be different," response bias is no longer controlled.

A further slight change in instructions can change the test from a triangle to a 3-AFC. An untrained experimenter might do this and alter the judges' cognitive strategy from 'comparison of distances' to 'skimming'. It is also theoretically possible that a judge might start out in a triangle test, correctly using a 'comparison of distances' strategy and then begin to notice the nature of the difference and change to a 'skimming' strategy (3-AFC). This would give a consequent improvement in performance. Such an improvement in performance at an inopportune time could give highly misleading results. The ASTM *Manual on Sensory Testing Methods* (1968) recommends the use of just the three triadic sequences of the triangle test for which the stronger stimulus is odd; this would seem to be inviting strategy change. Frijters *et al.* (1982) while trying to explain the incidental learning effect of McBride and Laing (1979), gave repeated triangles to see whether there might be a strategy change. They concluded that the performance obtained on their triangles did not indicate any strategy change. However, current work in our laboratory, where subjects think aloud during triangle testing, indicates that judges continually hypothesise about the nature of the difference, attempting to change to a 3-AFC. What can prevent them doing this is their lack of sensitivity, making it difficult for them to confirm their hypotheses and so adopt the 3-AFC 'skimming' strategy.

Interviews with judges have indicated that judges do not always use the strategy indicated by the instructions. Some judges use neither 'skimming' nor 'comparison of differences' but entirely novel strategies (Tedja *et al.*, 1994). It is naive to assume that a judge will automatically adopt the strategy suggested by the instructions. A strategy may need to be stabilised by experimental manipulation. For example, a comparison of differences strategy might be maintained by constantly changing the stimuli in the test, so that the judge cannot form expectations about the nature of the stimuli and so change to 'skimming'. 'Distractor' triangles could be placed randomly among the 'test' triangles to achieve this.

Is there a test that is sensitive and not so liable to strategy change as the triangle test? Both Thurstonian modelling and S.S.A. predict the directional paired comparison to be the most sensitive test with a powerful and stable skimming strategy. The usual objection to a paired comparison test is that it cannot be used, unless the nature of the difference can be described to the judge. Yet, this is not so. If, after 'warm-up' (O'Mahony *et al.*, 1988), the judge is asked to describe the nature of the difference, the forced-choice question can be posed in terms of the judge's description. Such an approach is called the warmed-up paired comparison (Thieme & O'Mahony, 1970); it is a simple procedure, yet only effective if the foods are susceptible to warm-up.

In summary, factors like cognitive strategy change and sequence effects, if not understood nor controlled, can render the triangle test seemingly variable in its results. A slight change in the instructions could cause strategy change or destroy the control of response bias. The triangle test is not a stable old workhorse like the paired comparison; it is fragile and prone to change, and it has the potential to produce confusing results when used by people who have not had an adequate background in sensory science.

REFERENCES

A. S. T. M. (1968). Forced choice methods. In *Manual on Sensory Testing Methods, Special Technical Publ. No. 434.* A. S. T. M., Philadelphia, pp. 24–49.

Berg, H. W., Filipello, F., Hinreiner, E. & Webb, A. D. (1955). Evaluation of thresholds and minimum difference concentrations for various constituents of wines. I. Water solutions of pure substances. *Food Technology*, **9**, (Jan), 23–6.

Byer, A. J. & Abrams, D. (1953). A comparison of the triangular and two-sample taste-test methods. *Food Technology*, **7** (April), 185–7.

David, H. A. & Trivedi, M. C. (1962). Pair, triangle and duo-trio tests. Technical Report No. 55, Dept. Statistics, Virginia Polytechnic Inst., Blacksburg, Virginia, USA.

Elliott, P. B. (1964). Tables of d'. In *Signal Detection and Recognition by Human Observers*, ed. J. A. Swets. John Wiley, New York, pp. 651–84.

Ennis, D. M. (1993). The power of sensory discrimination methods. *Journal of Sensory Studies*, **8**, 353–70.

Ennis, D. M. & Mullen, K. (1986). A multivariate model for discrimination methods. *J. Math. Psychol.*, **30**, 206–19.

Ennis, D. M. & Mullen, K. (1992). Probabilistic psychophysics with noisy stimuli. *Math. Soc. Sci.*, **23**, 221–34.

Ennis, D. M. & O'Mahony, M. Probabilistic models for sequential taste effects in triadic choice. *J. Exp. Psychol.*, in press.

Filipello, F. (1956). A critical comparison of the two-sample and triangular binomial designs. *Food Research*, **21**, 235–41.

Frijters, J. E. R. (1977). The effect of duration of intervals between olfactory stimuli in the triangular method. *Chem. Sens. Flav.*, **2**, 301–11.

Frijters, J. E. R. (1979). The paradox of discriminatory nondiscriminators resolved. *Chem. Senses*, **4**, 355–8.

Frijters, J. E. R. (1981). An olfactory investigation of the compatibility of oddity instructions with the design of a 3-AFC signal detection task. *Acta Psychol.*, **49**, 1–16.

Frijters, J. E. R. (1982). Expanded tables for conversion of a proportion of correct responses (P_c) to the measure of sensory difference (d') for the triangular method and the 3-alternative forced choice procedure. *J. Food Sci.*, **47**, 139–43.

Frijters, J. E. R., Blauw, Y. H. & Vermaat, S. H. (1982). Incidental training in the triangular method. *Chem. Senses*, **7**, 63–9.

Frijters, J. E. R., Kooistra, A. & Vereijken, P. F. G. (1980). Tables of d' for the triangular method and the 3-AFC signal detection procedure. *Percept. Psychophys.*, **27**, 176–8.

Geelhoed, E. N., MacRae, A. W. & Ennis, D. M. (1994). Preference gives more consistent judgments than oddity only if the task can be modeled as forced choice. *Percept. Psychophys.*, **55**, 473–7.

Green, D. M. & Swets, J. A. (1966). *Signal Detection-theory and Psychophysics.* John Wiley, New York.

Gridgeman, N. T. (1956). Group size in taste sorting trials. *Food Research*, **21**, 534–9.

Gridgeman, N. T. (1970). A re-examination of the two-stage triangle test for the perception of sensory differences. *J. Food Sci.*, **35**, 87–91.

Grim, A. C. & Goldblith, S. A. (1965). Some observed discrepancies in application of the triangle test to evaluation of irradiated whole egg magma. *Food Technology*, **19**, 146.

Hacker, M. J. & Ratcliff, R. (1979). A revised table of d' for M-alternative forced choice. *Percept. Psychophys.*, **26**, 168–70.

Halpern, B. P. (1986). What to control in studies of taste. In *Clinical Measurement of Taste and Smell*, eds H. L. Meiselman & R. S. Rivlin. MacMillan Publ. Co., New York, pp. 126–53.

Harries, J. M. (1956). Position bias in sensory assessments. *Food Technology*, **10** (Feb), 86–90.

Harrison, S. & Elder, L. W. (1950). Some applications of statistics to laboratory taste testing. *Food Technology*, **4** (Nov), 434–9.

Helm, E. & Trolle, B. (1946). Selection of a taste panel. *Wallerstein Lab. Communications*, **9**, 181–94.

Hopkins, J. W. (1954). Some observations on sensitivity and repeatability of triad taste difference tests. *Biometrics*, **10**, 521–30.

Lockhart, E. E. (1951). Binomial systems and organoleptic analysis. *Food Technology*, **5**, 428–31.

McBride, R. L. & Laing, D. G. (1979). Threshold determination by triangle testing: effects of judgemental procedure, positional bias and incidental training. *Chem. Senses Flav.*, **4**, 319–26.

MacRae, A. W. & Geelhoed, E. N. (1992). Preference can be more powerful than detection of oddity as a test of discriminability. *Percept. Psychophys.*, **51**, 179–81.

O'Mahony, M. (1979). Salt taste adaptation: The psychophysical effects of adapting solutions and residual stimuli from prior tastings on the taste of sodium chloride. *Perception*, **8**, 441–76.

O'Mahony, M. (1986). Sensory adaptation. *J. Sens. Stud.*, **1**, 237–58.

O'Mahony, M. (1989). Cognitive aspects of difference testing and descriptive analysis: Criterion variation and concept formation. In *Psychological Basis of Sensory-evaluation*, eds R. L. McBride & H. J. H. MacFie. Elsevier Applied Science, New York, pp. 117–39.

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. *J. Sens. Stud.*, **7**, 1–47.

O'Mahony, M. & Goldstein, L. R. (1986). Effectiveness of sensory difference tests: Sequential sensitivity analyses for liquid food stimuli. *J. Food Sci.*, **51**, 1550–3.

O'Mahony, M. & Goldstein, L. (1987). Tasting successive salt and water stimuli: the roles of adaptation, variability in physical signal strength, learning, supra- and subadapting signal delectability. *Chem. Senses*, **12**, 425–36.

O'Mahony, M. & Heintz, C. (1981). Direct magnitude estimation of salt taste intensity with continuous correction for salivary adaptation. *Chem. Senses*, **6**, 101–12.

O'Mahony, M., Masuoka, S. & Ishii, R. (1994). A theoretical note on difference tests; Models, paradoxes and cognitive strategies. *J. Sens. Stud.*, **9**, 247–72.

O'Mahony, M. & Odbert, N. (1985). A comparison of sensory difference testing procedures: Sequential sensitivity analysis and aspects of taste adaptation. *J. Food Sci.*, **50**, 1055–8.

O'Mahony, M., Thieme, U. & Goldstein, L. R. (1988). The warm-up effect as a means of increasing the discriminability of sensory difference tests. *J. Food Sci.*, **53**, 1848–50.

O'Mahony, M. & Wingate, P. (1974). The effect of interstimulus procedures on salt taste intensity functions. *Percept. Psychophys.*, **16**, 494–502.

Peryam, D. R. (1958). Sensory difference tests. *Food Technology*, **12** (May), 231–6.

Peryam, D. R. & Swartz, V. W. (1950). Measurement of sensory differences. *Food Technology*, **4** (Oct), 390–5.

Raffensberger, E. L. & Pilgrim, F. J. (1956). Knowledge of the stimulus variable as an aid in discrimination tests. *Food Technology*, **10** (Jan), 254–7.

Roessler, E. B., Pangborn, R. M., Sidel, J. L. & Stone, H. (1978). Expanded statistical tables for estimating significance in paired-difference, duo-trio and triangle tests. *J. Food Sci.*, **43**, 940–7.

Stillman, J. A. (1993). Response selection, sensitivity and taste-test performance. *Percept. Psychophys.*, **54**, 190–4.

Tedja, S., Nonaka, R., Ennis, D. M. & O'Mahony, M. (1994). Triadic discrimination testing: refinement of Thurstonian and Sequential Sensitivity Analysis approaches. *Chem. Senses*, **19**, 279–301.

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. *J. Sens. Stud.*, **5**, 159–76.

Thurstone, L. L. (1927*a*). Psychophysical analysis. *Am. J. Psychol.*, **38**, 368–89.

Thurstone, L. L. (1927*b*). A law of comparative judgement. *Psychol. Rev.*, **34**, 273–86.

Ura, S. (1960). Pair, triangle and duo-triangle test. *Reports of Statistical Application Research, Japanese Union of Scientists and Engineers*, **7**, 107–19.

Vié, A. & O'Mahony, M. (1989). Triangular difference testing: Refinements to sequential sensitivity analysis for predictions for individual triads. *J. Sens. Stud.*, **4**, 87–103.

Wasserman, A. E. & Talley, F. (1969). A sample bias in the evaluation of smoked frankfurters by the triangle test. *J. Food Sci.*, **34**, 99–100.