Available online at www.sciencedirect.com
ScienceDirect
Food Quality and Preference 18 (2007) 605–613
Food Quality and Preference
www.elsevier.com/locate/foodqual

Can the same-different test use a β-criterion as well as a τ-criterion?

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

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

Received 5 January 2005; received in revised form 20 February 2006; accepted 4 March 2006
Available online 16 January 2007

Abstract

Using low concentration NaCl and water stimuli, judges performed same-different tests and single stimulus discrimination tests. The data were subjected to a signal detection analysis. For single stimulus judgments, a β-criterion, dividing salt vs water is assumed for calculating d'. For the same-different method, a τ-criterion, a sensory yardstick designating the degree of difference required for a 'different' judgment, is assumed. ROC analysis indicated that for single stimulus judgments, a cognitive strategy involving a β-criterion was confirmed. Yet, ROC analysis for the same-different test indicated that prior or current use of a β-criterion carried over into the same-different test for some judges, giving a mixture of τ- and β-criteria.

Keywords: Same-different test; Signal detection; ROC curve; τ-criterion and β-criterion; Cognitive strategies; Decision rules

1. Introduction

Sensory difference tests are used for determining whether judges can discriminate between two foods which are so similar that they can be described as confusable. Such tests are used for quality assurance, ingredient specification, product development, and studies of the effects of processing change, packaging change and storage. Sometimes they are used analytically with trained panels and such tests then come under the general heading of what has been called Sensory Evaluation I (O'Mahony, 1995a). The complementary test is whether consumers can discriminate between foods under normal conditions of consumption, because differences that have been detected by a trained sensory panel in Sensory Evaluation I may not be detected by consumers under normal conditions of consumption. Such testing with untrained consumers then comes under the general heading of what has been called Sensory Evaluation II.

For the latter, it is generally desirable not to bias consumers by drawing attention to a particular attribute, so triangle and duo–trio tests are generally suitable. These methods lack power (Ennis, 1990, 1993) so suitable replacements would be desirable. One candidate is the same-different test, which can be more powerful than the duo–trio or triangle methods if used in a particular way (Ennis, 2004).

Each trial of the same-different discrimination test involves two samples, called the reference and comparison samples. The task requires a judge first to taste the reference sample and then taste the comparison sample, which may or may not be the same as the reference. The judge must then report whether the 'comparison' is the same as or different from the reference. This judgment has inherent response bias, so merely computing the proportion of correct responses does not give a representative measure of discrimination. However, a suitable signal detection/Thurstonian analysis can circumvent response bias and provide *R*-Index values (Cubero, Avancini de Almeida, & O'Mahony, 1995) or more fundamental d' values (O'Mahony & Rousseau, 2002).

The same-different method, unlike the duo–trio and triangle methods, does not have a standard form. For the two

* Corresponding author. Tel.: +1 530 752 6389; fax: +1 530 756 7320.
E-mail address: maomahony@ucdavis.edu (M. O'Mahony).

0950-3293/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.foodqual.2006.03.024

stimuli (W and S), there are four possible orders of presentation (WW, SS, WS, SW). For the short version of the test, a test is regarded as the presentation of just one of the four possible pairs. For the long version, a test consists of the presentation of two pairs, one pair the same (WW or SS) and one pair different (WS or SW). With this test, the judge is unaware that one pair is the same and the other different. As far as the judge is concerned, he responds to each pair as if he were performing two short version tests. It is this long version of the test that modeling has indicated has more power than the duo–trio or triangle methods (Ennis, 2004). Rousseau, Meyer, and O'Mahony (1998), using yoghurt stimuli, confirmed that the long version (but not the short version) of the same-different test was more powerful than the triangle method. The difference in power was only slight owing to relatively large d' values. More published confirmations would be desirable.

Some authors have used the short version (Lau, O'Mahony, & Rousseau, 2004; Rousseau & O'Mahony, 2001; Stillman & Irwin, 1995) while others used the long version (Rousseau & O'Mahony, 2000; Rousseau, Stroh, & O'Mahony, 2002) or both (Rousseau et al., 1998; Rousseau, Rogeaux, & O'Mahony, 1999). Authors have also used a modified method in which the reference is always the same stimulus (Avancini de Almeida, Cubero, & O'Mahony, 1999; Cubero et al., 1995; Delwiche & O'Mahony, 1996).

In the present study, using short-version same-different tests d' values were computed. A d' value is a measure of 'effect size' (Clark-Carter, 2003). It is a fundamental measure of the perceptual difference between two stimuli, measured in units of the perceptual variation of a single stimulus. An engineer might call it a signal-to-noise ratio. Its computation involves assumptions. Yet, if the assumptions are correct, a d' value should be independent of the method used to measure it. In this way, it is a fundamental measure. This is important for sensory evaluation because difference test measures have not been comparable between tests. The proportion of tests performed correctly with the duo–trio method cannot be compared to the proportion performed correctly for the triangle method, because their chance probabilities are different. Yet, their d' values can be compared.

What are the assumptions required of computing d' values? The first assumption is a convenience; it is that the sensory distributions of the two stimuli (W and S) are both normal and have equal variance. The equal variance assumption would seem logical when applied to confusable stimuli; if they are so similar that they can be confused, it would be no surprise that their variances would be the same. This assumption can easily be checked and experiments appear to support it (e.g. Hautus & Irwin, 1995; O'Mahony, 1972c). A more demanding assumption made to enable the estimation of d' is that of the nature of the cognitive strategy adopted by the judge. This always needs confirmation. For example, the computation of d' from the triangle test assumes the adoption of a 'comparison of distances' cognitive strategy, while that for the 3-AFC assumes a 'skimming' strategy (Ennis, 1993; O'Mahony, 1995b; O'Mahony, Masuoka, & Ishii, 1994). These assumptions needed confirming. Should judges perform both 3-AFC and triangle tests, then according to the assumptions, they should perform a greater proportion of 3-AFCs correctly, but the computation of the d' values, taking into account the appropriate cognitive strategies, should give the same results. This was confirmed by Tedja, Nonaka, Ennis, and O'Mahony (1994). A more sophisticated approach is to fit the various models to the data. For example, Irwin, Hautus and co-workers (Irwin, Hautus, & Stillman, 1992; Irwin, Stillman, Hautus, & Huddleston, 1993; Hautus & Irwin, 1995) fitted models that assumed different cognitive strategies to ROC curves, that they obtained using different experimental procedures, and determined which ones fitted the best.

In summary, a d' value for the same judge and the same pair of stimuli should be the same, regardless of the measurement method used, as long as the computation makes the correct assumptions. The most important assumption is the cognitive strategy. The assumptions allow the computation to circumvent the problems created by the differences in the experimental methods. It is because of this that investigation into the cognitive strategies associated with various test methods is important for sensory evaluation.

Brown (1974), when he developed his *R*-Index, wished for an index that was free of assumptions. It is equivalent to the index $P(A)$, the proportion of area underneath an ROC curve formed by connecting the points with straight lines (Green & Swets, 1966). It has been reviewed by O'Mahony (1992). The computation of $P(A)$ (*R*-Index) does not use assumptions to circumvent the differences between the experimental methods used to measure it. Accordingly, it is prone to vary with experimental method. For example, $P(A)$ for the same-different test will vary with the cognitive strategy used by the judge (see below). Also, it was shown that an *R*-Index obtained by ranking is higher than one obtained from rating data, because of the forced-choice nature of ranking (Ishii, Vié, & O'Mahony, 1992; O'Mahony, Garske, & Klapman, 1980). Thus, the *R*-Index or $P(A)$ can be seen as a measure of 'performance' rather than a fundamental measure of difference.

To explore the cognitive mechanisms associated with the same-different test, it is important to consider response biases and the criteria associated with difference tests. Consider a judge being given a set of confusable stimuli (W and S) and being required to report their identities (say "W" or "S"). This procedure has been called the yes/no task (Green & Swets, 1966). As the stimuli are confusable, the decision as to whether a stimulus is 'W' or 'S' can be difficult to make. His response will be the result of how well his receptors can distinguish between the two sensory signals elicited by 'W' and 'S' and also where he 'draws the line' between the sensations he would categorize as coming from 'W' and those he would categorize as coming from 'S' (Green & Swets, 1966; O'Mahony, 1992, 1995b).

Depending on where he 'draws his line', he may be biased and more willing to categorize his sensations as 'W' or biased towards categorizing his sensations as 'S': hence the term 'response bias'. The 'line' has a technical name. In food science, it is called the β-criterion (Rousseau, 2001; Rousseau et al., 1998). The cognitive strategy that uses this criterion has been called the β-strategy (Rousseau, 2001). Data from such an experiment can be used to produce ROC curves and to estimate d' values (Green & Swets, 1966). If a β-strategy were to be used in a same-different test, it would require the judge to identify each stimulus in the pair independently and then decide whether they fell on the same side or on different sides of the β-criterion line. In psychology, this strategy has been called the 'independent observation model' (Macmillan & Creelman, 1991) and the 'optimal' strategy or decision rule (Irwin & Francis, 1995; Noreen, 1981). The 'optimality' of the β-strategy relates to the fact that $P(A)$ (R-Index), the proportion of area below the ROC curve for a given value of d', turns out to be larger for this strategy than for the τ-strategy.

To digress briefly, it can be seen that psychology and sensory food science use a different set of technical terms and symbols. Because it is sometimes necessary for the food scientist to explore the psychological literature, it is as well to be aware of these differences. For example, the psychologists Green and Swets (1966) denote the β-criterion by the symbol 'k'. To add to the confusion, they use the symbol 'β' to denote something completely different: the likelihood ratio at the criterion point. One crosses interdisciplinary boundaries with care.

In a same-different test, a judge has two possible cognitive strategies at his disposal (Hautus & Irwin, 1995). Besides the β-strategy, the judges can use a second strategy. This involves the use of a τ-criterion (Rousseau, 2001; Rousseau et al., 1998). A τ-criterion is concerned with how different two stimuli need to be, to be reported as 'different'. It can be visualized as a sensory yardstick. If sensations elicited by the two stimuli in the same-different test are more different than the yardstick, the stimuli will be reported as different; if not, they will be reported as the same (Irwin & Francis, 1995; Irwin et al., 1993; Rousseau, 2001; Rousseau et al., 1998). In psychology, the τ-criterion has been called a k-criterion (Macmillan & Creelman, 1991; Macmillan, Kaplan, & Creelman, 1977). The cognitive strategy that utilizes the τ-criterion will here be called the τ-strategy. In psychology, it has been called the 'differencing model' (Macmillan & Creelman, 1991) or the 'sensory difference decision rule' (Noreen, 1981).

For a standard yes/no task, in which a judge is presented with either S or W and must indicate which sample was presented, the β-strategy is the only available strategy that can be adopted. For this case, d' can be calculated from the standard yes/no ROC curve (Green & Swets, 1966). If the sensation distributions are normal with equal variance, the ROC curve will be symmetrical about the negative diagonal and a z-plot ROC will be a straight line with a slope of unity. The regular ROC curve will lose its symmetry and the slope of the z-plot will deviate from unity if the variances of the two distributions are not equal; the slope will be equal to the ratio of the standard deviations of the two distributions.

This standard ROC analysis used for the yes/no task cannot be applied to the same-different task. A different approach is required to compute d'. There is one approach if a β-strategy is used and another if the τ-strategy is adopted. If a β-strategy is used for the same-different test, the ROC will be symmetrical, as is the ROC for the standard yes/no method with equal variance. However, the same-different ROC will not be exactly the same shape as that for the standard yes/no model. It will be symmetrical, yet it will have a higher proportion of area, $P(A)$, under the curve. If the two curves were to be overlaid, the same-different ROC would be seen to bulge out further than the yes/no curve. Yet, the two curves would meet where they intersect the negative diagonal and, of course, would also meet at the ends of the curve. So the same-different curve can be described as rising up more quickly and then flattening out to meet the lower yes/no curve at the negative diagonal. Yet, even though the $P(A)$ values would be different, the d' values, taking into account the different cognitive strategies, should be the same.

If a τ-strategy is used, the ROC will be asymmetrical as is the ROC for the standard yes/no model with unequal variance. Again, the two ROCs will not be the same shape, however the differences are more complex than those for the β-strategy ROC given above. Again, for the same-different model, the proportion of area under the curve will be greater than for the yes/no task with unequal variance, yielding higher $P(A)$ (R-Index) values.

To summarize, ROC curves for same-different method using the β-strategy and for the yes/no method (equal variances) using the β-strategy are both symmetrical. Yet $P(A)$ for the same-different method using the β-strategy is greater. The ROC curves for the same-different method using the τ-strategy and for the yes/no method using the β-strategy (unequal variances) are both asymmetrical but $P(A)$ is greater for the same-different method. However, for all these procedures the computed d' values should be the same, if the computation takes account of the appropriate cognitive strategy. In addition, Hautus and Irwin (1995) indicated that $P(A)$ values were greater for the yes/no method (β-criterion) than for the same different method (τ-criterion); again computed d' values will be the same if the appropriate cognitive strategy is assumed.

Because it is important to know the cognitive strategy being used in a difference test to be able to compute d', it is important to investigate such strategies. One approach to determining the cognitive strategy is simply to interview the judge (Tedja et al., 1994) or to require the judge to 'think aloud' (Wong, 1997). Because judges may not always be aware of their cognitive strategy, a second approach is to examine the ROC curve obtained for the judge. The present study relied on this latter approach although the occasional judge would 'think aloud' and was not discouraged.

The fact that the ROCs for the same-different β-strategy and the standard yes/no task with equal variances are both symmetrical, and those for the same-different τ-strategy and the standard yes/no task with unequal variances are both asymmetrical, leads to a simplified ROC analysis to determine whether a τ-criterion is being used for the same-different task. Simply fit the standard yes/no ROC to the same-different data to determine whether or not the ROC curve is symmetrical. If the variances of the two sensory distributions can be assumed to be the same, the standard yes/no ROC curve will be asymmetric for a τ-strategy (Hautus & Irwin, 1995; Irwin et al., 1993). Also, the slope of the *z*-plot ROC will be greater than unity. This shortcut method is only useful for determining the cognitive strategy being used. It will not provide a legitimate estimate of d'. The computation of the d' value is more complex. It requires fitting the appropriate same-different model of the ROC to the data. Alternatively, for the τ-strategy, d' can be estimated without undertaking an ROC analysis, by using Ennis's method of computation given in O'Mahony and Rousseau (2002).

In psychology, there has been considerable discussion regarding cognitive strategies or decision rules involved in the same-different method (for example, Dai, Versfeld, & Green, 1996; Irwin & Hautus, 1997; Noreen, 1981; Sorkin, 1962). Kaplan, Macmillan, and Creelman (1978) provided tables of d' for the same-different test. Irwin and Francis (1995) noted that different cognitive strategies were adopted for visual stimuli, depending on the complexity of the stimuli. A set of judges used a β-strategy for complex stimuli (kanji: Japanese system of writing using Chinese characters) while with what appeared to be a separate group of judges, a τ-strategy was adopted by two out of three judges for more simple stimuli (colors). They also studied same-different tests, where the stimuli were judged as conceptually the same or different (natural vs manufactured items), rather than physically so (Francis & Irwin, 1995; Irwin & Francis, 1995); their results supported a β-strategy.

Yet, it is in the work of Irwin, Hautus, and their co-workers, who considered taste and food stimuli, that the considerations of cognitive strategy become more relevant to sensory evaluation. Irwin et al. (1992) reviewed ROC curves and some of the drawbacks of *R*-Index measurement for methods that induced a β-criterion. Irwin et al. (1993) demonstrated how ROC curves, derived from same-different tests for orange drinks, were best fitted assuming a cognitive strategy that used a τ-criterion. The same result was obtained by Stillman and Irwin (1995) using a raspberry flavored drink. These data were supported by same-different experiments with auditory stimuli (Hautus, Irwin, & Sutherland, 1994). From these studies, it would seem that in the same-different test, judges tend spontaneously to adopt a τ-cognitive strategy. Irwin, Hautus and co-workers went on to examine bias and the interpretation of areas under ROC curves for the same-different test (Irwin, Hautus, & Butcher, 1999; Irwin, Hautus, & Francis, 2001). However, the most relevant paper to the present study is that of Hautus and Irwin (1995).

Hautus and Irwin used the signal detection rating procedure to determine how well judges could distinguish between milks of different fat content. In a yes/no task, a random order of milk stimuli was tasted and judges had to report which milk they tasted and rate the sureness of their responses on a six-point category scale. Values of d' were calculated and symmetrical ROC curves obtained (using the standard yes/no model), indicating that the two sensory distributions had equal variance. In a second experiment, the same stimuli were discriminated using a same-different test, again with sureness ratings on a six-point scale. The ROC curves obtained were asymmetric and the d' values calculated assuming a τ-criterion, agreed with those in the first experiment. The d' values were fairly close to threshold, so it was not possible to tell whether an analysis assuming a τ criterion would have fitted the data better. However, from past research, it would seem unlikely.

In Experiment I for the present study, the goal was to determine whether it was possible for judges to use a β-strategy for a same-different taste test. Accordingly, judges performed short-version same-different tests under two protocols. For one protocol, the experimental conditions were set up to favor a τ-strategy while for a second protocol they favored a β-strategy. ROC curves were examined to determine which strategy was actually used for each protocol. Should the β- and τ-strategies be used in their appropriate protocols, then it may be expected that computed d' values computed from both protocols and also from some additional 2-AFC tests should correspond.

2. Experiment I

2.1. Materials and methods

2.1.1. Judges

Eleven judges (3 M, 8 F; age range 21–62 years), students, staff and friends at UC Davis, participated in the experiment. Judges were required to fast, except for water, for at least 1 h prior to testing. Five had participated in taste psychophysical experiments beforehand, six had not.

2.1.2. Stimuli

Stimuli consisted of low concentration NaCl solutions (0.5–5.0 mM, depending on the judge's sensitivity) to be discriminated from purified water. The NaCl solutions (S) were prepared by dissolving reagent grade NaCl (Mallinckrot Inc., Paris, KY) in Milli-Q purified water. The Milli-Q purified water was deionized water fed into a Milli-Q system involving ion exchange and activated charcoal (Millipore Corp., Bedford, MA). The resulting purified water had a specific conductivity of $<10^{-6}$ mho/cm and a surface tension $\geqslant 71$ dynes/cm. The purified water was used as the water stimulus (W).

Stimuli were dispensed in 10 ml aliquots using both Repipet Adjustable Dispensers (Labindustries Inc., Berkeley, CA)

and Oxford Adjustable Dispensers (Lancer, St Louis, MO) in plastic cups (1 oz portion cups, Solo Cup Co., Urbana, IL). All stimuli were served at constant room temperature (21–24 °C), on white plastic cutting trays. Stimulus concentrations ranged 0.5–5.0 mM (0.5–3.0 mM, two judges; 1.0–3.0 mM, seven judges; 3.0–5.0 mM, two judges).

2.1.3. Procedure

Each judge performed 96 same-different tests under each of two protocols. In the β-protocol, conditions were arranged to encourage the use of a β-criterion. In the τ-protocol, a τ-criterion was encouraged. The tests were performed over four separate sessions on separate days (24 tests under each protocol per session, total = 96 per protocol).

For what will be called the β-protocol, judges first performed a warm-up procedure (Dacremont, Sauvageot, & Duyen, 2000; O'Mahony, Thieme, & Goldstein, 1988; Pfaffmann, 1954; Thieme & O'Mahony, 1990). The warm-up consisted of tasting alternately water and salt stimuli, so the judge could discover the sensory signals that denoted each one. In other words, the judge was establishing a β-criterion, differentiating between water and salt. Each judge tasted at least five of each stimulus and more if desired. After the warm-up, judges then performed 24 same-different tests which had been modified to promote the use of a β-criterion. The judge, when presented with the pair of stimuli, was required to report whether each stimulus was water or salt. If both were reported as water or as salt, the test response was scored as "same". If one of the stimuli was reported as salt and the other as water, the test response was scored as "different". Judges were also required to say whether they were sure or unsure of their pair of judgments.

For what will be called the τ-protocol, a modified warm-up procedure was used. Judges were presented with two pairs of stimuli. The first pair consisted of two water stimuli and the second pair consisted of water followed by salt (W–W, W–S). Judges tasted at least five of each of these pairs and more if desired. Next, two further pairs (S–S, S–W) were tasted in the same way. The goal of this warm-up was for judges to discover the signals indicating same and different stimulus pairs and thereby to establish a τ-criterion, indicating the degree of difference required for a "different" response. After this modified warm-up, judges then performed 24 regular same-different tests. Again, judgements of "sure"/"not sure" were added. The judges were discouraged from identifying the stimuli and instructed only to pay attention to whether they felt the stimuli were the same or different. Subjective responses indicated that judges could perform according to each protocol; four judges who reported difficulty with these tasks were eliminated.

After establishing rapport, and taking demographic details, the experimenter instructed the judge to take at least six purified water mouthrinses to clean the mouth. Judges then performed the warm-up for the specific protocol and the same-different tests without any interstimulus rinsing. After a further six mouthrinses, the warm-up procedure and same-different tests were performed for the other protocol.

Also included in each session were twelve 2-AFC tests. Before these tests, judges took six mouthrinses, and performed a warm-up with at least five pairs of water and salt stimuli. No further mouthrinses were taken after the initial six. The 2-AFC tests were performed either at the beginning or the end of an experimental session. There were thus, four possible orders of presentation for the β-protocol, τ-protocol and 2-AFC protocol within an experimental session. These were β/τ/2-AFC, 2-AFC/τ/β, τ/β/2-AFC, and 2-AFC/β/τ. The four orders of presentation were all used for each judge over the four experimental sessions. The order of presentation of stimuli within a given test, was chosen randomly for one judge and the reverse order was employed for the next judge. For the third judge, a separate random order was chosen, and so on. Subjects responded verbally. Experimental session lengths ranged 15–43 min.

Prior to the first experimental session, judges had a training session. Judge sensitivity was determined using 2-AFC tests. Judges who could not discriminate between purified water and 5 mM NaCl were eliminated. Judges practiced using the β and τ-protocols and those who experienced difficulty using the two cognitive tasks were eliminated based on their subjective reports. From the results of this practice session, it was decided that 5 mM NaCl should be used for the same-different tests in the main experiment. However, it was necessary to use a lower concentration (3 mM) for pairs where NaCl was tasted after purified water. This was because of sensitization to NaCl caused by a lowering in the adaptation level by the water stimulus (Bartoshuk, 1968, 1974, 1978; Halpern, 1986; McBurney & Pfaffmann, 1963; O'Mahony, 1972a, 1972b, 1979; O'Mahony & Godman, 1974). Thus, the session used 5 mM NaCl, 3 mM NaCl and water. If a ceiling effect was encountered or a more sensitive judge was tested, the concentrations were reduced to 3 and 1 mM NaCl. A further reduction to 1 and 0.5 mM NaCl was found necessary for two judges. Nine judges started with 3 and 1 mM, while two started at 5 and 3 mM. Because the goal of the experiment was only to compare performance on the same-different test under the β- and τ-protocols, the variations in concentration for these judges did not invalidate the experiment because exactly the same concentrations were used an equal number of times under each protocol.

2.2. Results and discussion

Mean d' values, computed from the 11 judges tested in Experiment I, are shown in Table 1 using a variety of models. The values were computed using the IFPrograms software (Institute for Perception, Richmond, Virginia). Significant differences between these means were computed

Table 1
Mean d' values computed using various analyses from the results of Experiment I ($N = 11$)

	Same-different test	
	τ-Protocol	β-Protocol
From 2-AFC method	From same-different judgments using computation that assumes τ-criteria	From salt vs water judgments using computation that assumes β-criteria
1.86[a]	1.82[a]	1.54[a]

[a] Means were not significantly different ($p > 0.05$).

using ANOVA and LSD tests ($p < 0.05$). The first (1.86) was computed from the 2-AFC tests and can also be derived from Tables (Ennis, 1993). The second (1.82) value was derived from the same-different test performed in the τ-protocol, using a computation (degree of difference program, IFPrograms) that assumed a τ-criterion (O'Mahony & Rousseau, 2002). The third mean (1.54) was derived from salt vs water judgments from the β-protocol using a computation (scale program, IFPrograms) that assumed a β-criterion (Kim, Ennis & O'Mahony).

To investigate whether τ- or β-strategies were used in the same-different tests, ROC curves were fitted to the data in the different/same matrices using maximum-likelihood estimation (Hautus et al., 1994). The data were pooled across judges. This has drawbacks for estimating sensitivity (Macmillan & Creelman, 2005). However, if sensitivity is not the main area of interest, then pooling data can give a stricter test of the models under investigation since the variability of each point on the ROC curve can be dramatically reduced. The pooled data for the τ and β protocols are given in Table 2. The table indicates the best fitting value of d', the goodness-of-fit statistic (χ^2), and the probability that the data arose from the model fitted, given that the model was correct (p). Smaller values of χ^2, and larger values of p, indicate a relatively better fit. It can be seen from the table for both χ^2 and p, the results suggest that the best overall fit in both protocols was the τ-strategy.

Considering the three d' values from Table 1, the fact that they were not significantly different would confirm signal detection/Thurstonian theory. Each value had been computed taking into account the appropriate cognitive strategy. Yet, the value for the β-protocol was rather low. However, the ROC analyses indicated that a τ-strategy rather than a β-strategy was being used for this protocol. In this case, because $P(A)$ is smaller for the τ-strategy, a computation assuming a β-strategy would underestimate d'.

There are additional possibilities. Because the sureness judgments for the τ-protocol were given for pairs of stimuli rather than for each individual stimulus some boundary variance may have been introduced. Boundary variance is variance introduced because judges vary in their β-criteria (boundaries) between what should be reported as "water" or "salt" and their judgements of "sure" and "unsure". Such added variance would depress the value of d'.

Thus, from the pooled data, it may be concluded that this experiment failed to demonstrate that a β-strategy would be adopted by judges in a condition favorable to its adoption (τ-protocol). Yet, this conclusion is not so clear when data from individual judges are examined. For the τ-protocol, the data computed for each individual judge indicated that six judges had a better fit with the β-strategy while only five had a better fit with the τ-strategy. Surprisingly, for the τ protocol, only two had a better fit for the β-strategy while nine had a better fit for the τ-strategy. There are various explanations for this mix of strategies. It may be that the four category sureness scale used to generate the data did not give sufficiently accurate ROCs. It may be that judges carried over their strategies from one protocol to another. It might also be that some of the data could be the result of the intrusion of a third unexpected strategy. Unexpected strategies have been reported before (Tedja et al., 1994). However, it must be stressed that the results for individual judges must be interpreted with care. They are more prone to sampling error compared with the pooled results.

Notwithstanding, the fact that some judges might have used a β-strategy requires more attention. Accordingly, the possibility of the use of a β-strategy was investigated further in Experiment II.

Table 2
Results of ROC analyses for pooled data for τ and β protocols in Experiment I

Protocol	Fitted cognitive strategy					
	τ-Strategy			β-Strategy		
	d'	χ^2	p^*	d'	χ^2	p
τ	1.72	**2.43**	**0.296**	1.40	7.74	0.021
β	2.38	**1.74**	**0.419**	1.93	8.51	0.014

Small χ^2 and large p signifies a good fit. Bold and underlined χ^2 and p values indicate best fitted strategy for ROC curve.

[*] p = probability that data arose from the model fitted given that the model is correct.

3. Experiment II

The goal of this experiment was to use an improved experimental design to determine whether a prior set of single stimulus judgments, requiring a β-strategy, could affect the choice of strategy for a subsequent same-different test. To enable better fitting of ROC curves than in Experiment I, judges were required to perform more tests and responses were given on a six-point rather than a four-point scale. Also, to check that prior judgments of single stimuli did conform to the β-strategy, independent sureness judgments were made for each stimulus to allow the construction of ROC curves. These curves which were not available from the first experiment because of the sureness judgment regime used, were available for comparison with same-different ROCs.

3.1. Materials and methods

3.1.1. Judges

Four judges (4 F; age range 22–44 years) who had participated in Experiment I were available to return for more intensive re-testing.

3.1.2. Stimuli

The stimuli were the same as in Experiment I with NaCl concentrations adjusted to produce d' values in the range 1.8–2.5. This avoided ROC curves being too close to the positive diagonal, where it is not easy to distinguish between ROC curves generated by β- and τ-criteria. Stimulus concentrations ranged 1.0–5.0 mM (1.0–3.0 mM, one judge; 3.0–5.0 mM, three judges). Because the goal was to determine the shape of the ROC curves and not make estimates of d', combining data over sessions with slightly different signal strengths did not invalidate the data.

3.1.3. Procedure

Each judge performed 24 tests per experimental session. In each experimental session a test consisted of tasting a pair of stimuli. The judge was required to rate the first stimulus as being 'salt' or 'water' using three levels of sureness: "sure" vs "not sure" vs "I do not know but I will guess" resulting in a six-point category scale. This is the signal detection rating procedure described by Green and Swets (1966). The judge was then required to rate the second stimulus in the same way. Finally, she was required to judge whether the two stimuli were the same or different, using the three levels of sureness for this same-different judgment. All four possible pairs were used (WS, SW, WW, SS).

To check whether the judge's daily variation in sensitivity was in the range that was advantageous for fitting same-different ROC curves ($d' = 1.80$–2.50), a set of six 2-AFC tests was performed prior to and after the end of the testing. Judges began by taking at least six mouthrinses. They then performed a warm-up as in Experiment I for the τ-protocol. Immediately after this, they performed six 2-AFC tests. They were then offered the option of a further warm-up if desired before beginning the 24 tests. After the 24 sets of same-different tests, judges were given a further six 2-AFCs without prior warm-up just to double-check the sensitivity change after the tests. Judges performed eight sessions, giving a total of 192 same-different tests.

If the initial six 2-AFC tests indicated that the judge did not have the required sensitivity, a further six 2-AFCs were performed as a check before any decision was made about abandoning that experimental session. If the data analysis for a given session indicated that the judge's sensitivity was not in the specified range, the session was rescheduled and repeated. The number of abandoned sessions ranged 2–8 per judge. As the sensitivity of judges changed with practice, the NaCl stimulus concentrations were varied to keep the judges within the required sensitivity range. Experimental session lengths ranged 20–45 min. Generally, experimental sessions were performed on separate days although some judges chose to perform more than one session per day. Testing ranged over 7–11 days per judge. Other details of the procedure were as for Experiment I.

3.2. Results and discussion

ROC curves were fitted to data obtained from the first stimulus and the second stimulus and the same-different judgments. They were fitted using maximum-likelihood estimation (Hautus et al., 1994). The results are shown in Table 3. Considering the results for the single stimulus

Table 3
Results of ROC analyses for same-different test with single stimulus and same-different judgments in Experiment II

Judge	ROC analysis for single stimulus judgments (β-strategy)							
	First stimulus				Second stimulus			
	d'	Slope	χ^2	p	d'	Slope	χ^2	p
A	2.20	1.67	**4.14***	**0.246**	1.43	0.94	**1.43**	**0.698**
B	2.97	2.26	**5.55**	**0.136**	1.70	2.40	**5.64**	**0.130**
C	2.53	1.18	7.66	0.054	1.86	0.79	**3.00**	**0.392**
D	2.26	1.49	**5.35**	**0.148**	1.30	1.99	**0.21**	**0.977**
Pooled	2.30	1.69	21.8	<0.001	1.50	1.35	1.28	0.733
	ROC analysis for same-different judgments							
	τ-Strategy				β-Strategy			
	d'	χ^2	p		d'	χ^2	p	
A	2.29	**3.89****	**0.421**		1.80	8.79	0.066	
B	2.76	7.19	0.126		2.19	**3.15**	**0.532**	
C	2.47	6.39	0.172		1.96	**5.20**	**0.267**	
D	2.27	**3.12**	**0.537**		1.88	4.26	0.372	
Pooled	2.38	6.62	0.157		1.90	11.16	0.025	

* Statistic and probability for goodness of fit (β-strategy) for the single stimulus judgment in bold.
** Statistic and probability for best fitted strategy for same-different judgments for ROC curves in bold and underlined.

judgments, the standard β-strategy (yes/no) model was fitted. The fit of the model for the ROC curves was good in all cases except for the judge C with the first stimulus. This indicated that the judges were, at this point, decision making in terms of β-criteria. The slopes for both stimuli tended to be greater than unity, indicating a larger variance for the 'noise'; values less than unity were not significantly so. For all judges, d' for the first stimulus was higher than for the second stimulus. An examination of data indicated that this was mainly due to errors when both stimuli were salt. This is expected from earlier research on sequence effects and is predicted by the Conditional Stimulus and Sequential Sensitivity Analysis models of discrimination (Ennis & O'Mahony, 1995; O'Mahony & Goldstein, 1987; O'Mahony & Odbert, 1985; Tedja et al., 1994; Vié & O'Mahony, 1989). Thus, the first stimulus is the better representative of the use of a β-strategy for single stimulus judgments.

Considering the fit for the same-different judgments at the bottom of the table, the χ^2 and p values were consistent with the judges A and D using the τ-strategy while judges B and C used the β-strategy. It may be hypothesized that judges B and C carried over their β-decision making strategies into the same-different test while judges A and D did not. This would confirm the results from Experiment I where some judges used a τ-strategy while others used a β-strategy. This is interesting because, as outlined in the Introduction section, research generally indicates a τ-strategy for same-different judgments. The present research indicates that the choice of τ-strategy can be interfered with by prior β-strategy decision making. Further research should broaden the knowledge on strategy choice for the same-difference test.

Three judges (A, B, and C) were consistent in their choice of strategy for same-different judgments in Experiments I and II. The fourth inconsistent judge (D) did not fit either strategy well in Experiment I. Values of d' for the first single stimulus and the same-different test favoring the τ-strategy (judges A and D) showed good agreement. Yet, for judges B and C (β-strategy), agreement was not good, although sometimes complicated by low p-values.

In conclusion, although the same-different method with simple sensory stimuli has generally been shown to use a τ-criterion, the present research indicates that if judges were involved in prior or simultaneous decision making using a β-criterion, a β-decision rule or cognitive strategy may be adopted by some judges for the same-different test. This adds to the work of Irwin and Francis (1995) and Francis and Irwin (1995) who found that for conceptual same-different judgments and judgments of complex stimuli, a β-strategy best fitted ROC data.

One may speculate that judges, if exposed to the same or similar stimuli over a period of time, may begin to categorize stimuli and begin to use a β-criterion for making their same-different judgements. Also, it may be hypothesized that what applies to the same-different test in terms of cognitive strategies may also apply to the versions of the A-Not A method (ASTM, 1968; Peryam, 1958; Pfaffmann, 1954). These are all topics for further research.

References

ASTM (1968). *Manual on sensory testing methods STP 434*. Philadelphia: American Society for Testing and materials.

Avancini de Almeida, T. C., Cubero, E., & O'Mahony, M. (1999). Same-different discrimination tests with interstimulus delays up to one day. *Journal of Sensory Studies, 14*, 1–18.

Bartoshuk, L. M. (1968). Water taste in man. *Perception and Psychophysics, 3*, 69–72.

Bartoshuk, L. M. (1974). NaCl Thresholds in man: thresholds for water taste or NaCl taste? *Journal of Comparative and Physiological Psychology, 87*, 310–325.

Bartoshuk, L. M. (1978). The psychophysics of taste. *American Journal of Clinical Nutrition, 31*, 1068–1077.

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

Clark-Carter, D. (2003). Effect size: the missing piece of the jigsaw. *The Psychologist, 16*, 636–638.

Cubero, E., Avancini de Almeida, T. C., & O'Mahony, M. (1995). Cognitive aspects of difference testing: memory and interstimulus delay. *Journal of Sensory Studies, 10*, 307–324.

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.

Dai, H., Versfeld, N. J., & Green, D. M. (1996). The optimum decision rules in the same-different paradigm. *Perception and Psychophysics, 58*, 1–9.

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.

Ennis, D. M. (1990). Relative power of difference testing methods in sensory evaluation. *Food Technology, 44*, 114, 116, 117.

Ennis, D. M. (1993). The power of sensory discrimination methods. *Journal of Sensory Studies, 89*, 353–370.

Ennis, D. M. (2004). Personal communication.

Ennis, D. M., & O'Mahony, M. (1995). Probabilistic models for sequential taste effects in triadic choice. *Journal of Experimental Psychology, 21*, 1088–1097.

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 detection theory and psychophysics*. New York: John Wiley.

Halpern, B. P. (1986). What to control in studies of taste? In H. L. Meiselman & R. S. Rivlin (Eds.), *Clinical Measurement of Taste and Smell* (pp. 126–153). New York: Macmillan.

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

Hautus, M. J., Irwin, R. J., & Sutherland, S. (1994). Relativity of judgements about sound amplitude and the asymmetry of the same-different ROC. *Quarterly Journal of Experimental Psychology, 47A*, 1035–1045.

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.

Irwin, R. J., & Hautus, M. J. (1997). Likelihood-ratio decision strategy for independent observations in the same-different task: an approximation to the detection-theoretic model. *Perception and Psychophysics, 59*, 313–316.

Irwin, R. J., Hautus, M. J., & Butcher, J. C. (1999). An area theorem for the same-different experiment. *Perception and psychophysics, 61*, 766–769.

Irwin, R. J., Hautus, M. S., & Francis, M. A. (2001). Indices of response bias in the same-different experiment. *Perception and Psychophysics, 63*, 1091–1100.

Irwin, R. J., Hautus, M. J., & Stillman, J. A. (1992). Use of the receiver operating characteristic in the study of taste perception. *Journal of Sensory Studies, 7*, 291–314.

Irwin, R. J., Stillman, J. A., Hautus, M. J., & Huddleston, L. M. (1993). The measurement of taste discrimination with the same-different task: a detection-theory analysis. *Journal of Sensory Studies, 8*, 229–239.

Ishii, R., Vié, A., & O'Mahony, M. (1992). Sensory difference testing: ranking *R*-indices are greater than rating *R*-indices. *Journal of Sensory Studies, 7*, 57–61.

Kaplan, H. L., Macmillan, N. A., & Creelman, C. D. (1978). Tables of d' for variable-standard discrimination paradigms. *Behavior Research Methods and Instruments, 10*, 796–813.

Lau, S., O'Mahony, M., & Rousseau, B. (2004). Are three-sample tasks less sensitive than two-sample tasks? Memory effects in the testing of taste discrimination. *Perception and Psychophysics, 66*, 464–474.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: a user's guide*. New York: Cambridge University Press.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: a user's guide* (2nd ed.). New York: Cambridge University Press.

Macmillan, N. A., Kaplan, H. L., & Creelman, C. D. (1977). The psychophysics of categorical perception. *Psychological Review, 84*, 452–471.

McBurney, D. C., & Pfaffmann, C. (1963). Gustatory adaptation to saliva and sodium chloride. *Journal of Experimental Psychology, 65*, 523–529.

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. (1972a). The interstimulus interval for taste. 1: The efficiency of expectoration and mouth rinsing in clearing the mouth of salt residuals. *Perception, 1*, 209–215.

O'Mahony, M. (1972b). The interstimulus interval for taste. 2: Salt taste sensitivity drift and the effects on intensity scaling and threshold measurement. *Perception, 1*, 217–222.

O'Mahony, M. (1972c). Salt taste sensitivity: a single detection approach. *Perception, 1*, 439–464.

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*, 441–476.

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. (1995a). Sensory measurement in food science: fitting methods to goals. *Food Technology, 29*, 72–82.

O'Mahony, M. (1995b). Who told you the triangle test was simple? *Food Quality and Preference, 6*, 227–238.

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., & Godman, L. (1974). The effect of interstimulus procedures on salt taste thresholds. *Perception and Psychophysics, 16*, 459–465.

O'Mahony, M., & Goldstein, L. (1987). Tasting successive salt and water stimuli: the roles of adaptation, variability in physical signal strength, learning, supra- and subadaptating signal detectability. *Chemical Senses, 12*, 425–436.

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., & Odbert, N. (1985). A comparison of sensory difference testing procedures: sequential sensitivity analysis and aspects of taste adaptation. *Journal of Food Science, 50*, 1055–1058.

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*, 231–236.

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

Rousseau, B. (2001). The β-strategy: an alternative and powerful cognitive strategy when performing sensory discrimination tests. *Journal of Sensory Studies, 16*, 301–318.

Rousseau, B., Meyer, A., & O'Mahony, M. (1998). Power and sensitivity of the same-different test: comparison with triangle and duo–trio methods. *Journal of Sensory Studies, 13*, 149–173.

Rousseau, B., & O'Mahony, M. (2000). Investigation of the effect of within-trial retasting and comparison of the dual-pair, same-different and triangle paradigms. *Food Quality and Preference, 11*, 457–464.

Rousseau, B., & O'Mahony, M. (2001). Investigation of the dual-pair method as a possible alternative to the triangle and same-different tests. *Journal of Sensory Studies, 16*, 161–178.

Rousseau, B., Rogeaux, M., & O'Mahony, M. (1999). Mustard discrimination by same-different and triangle tests: aspects of irritation and τ criteria. *Food Quality and Preference, 10*, 173–184.

Rousseau, B., Stroh, S., & O'Mahony, M. (2002). Investigating more powerful discrimination tests with consumers: effects of memory and response bias. *Food Quality and Preference, 13*, 39–45.

Sorkin, R. D. (1962). Extensions of the theory of signal detectability to matching procedures in psychoacoustics. *Journal of the Acoustical Society of America, 34*, 1745–1751.

Stillman, J. A., & Irwin, R. J. (1995). Advantages of the same-different method over the triangular method for the measurement of taste discrimination. *Journal of Sensory Studies, 10*, 261–272.

Tedja, S., Nonaka, R., Ennis, D. M., & O'Mahony, M. (1994). Triadic discrimination testing: refinement of Thurstonian and Sequential Sensitivity Analysis approaches. *Chemical 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. *Journal of Sensory Studies, 5*, 159–176.

Vié, A., & O'Mahony, M. (1989). Triangular difference testing: refinements to sequential sensitivity analysis for predictions for individual triads. *Journal of Sensory Studies, 4*, 87–103.

Wong, D. (1997). *Cognitive strategies of the triangle difference test*. MS Thesis, University of California, Davis.