Food Quality and Preference 31 (2014) 49–55

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Triangle and tetrad protocols: Small sensory differences, resampling and consumer relevance

Rie Ishii [a], Michael O'Mahony [a], Benoît Rousseau [b,*]

[a] University of California, Davis, CA, USA
[b] The Institute for Perception, Richmond, VA, USA

ARTICLE INFO

Article history:
Received 27 September 2012
Received in revised form 9 July 2013
Accepted 14 July 2013
Available online 24 July 2013

Keywords:
Tetrad
Triangle
Discrimination testing
Power
Resampling

ABSTRACT

The objective of this research was twofold: first, the performance of the tetrad protocol was compared to that of the triangle test under conditions that could possibly lower its sensitivity, consequently resulting in the loss of its theoretical power advantage. Second, the same samples were compared with a preference test to investigate whether a no difference conclusion obtained with a discrimination test would consistently result in a non-significant preference (consumer relevance).

The investigation involved sensory differences that could be deemed small (d' values less than 1.0) as well as the comparison of resampling vs. no resampling conditions. 456 consumers performed tests using apple and orange juices for which slight sensory differences were created through dilution. In all conditions, the tetrad always exhibited a greater number of correct answers than the triangle, confirming its greater statistical power. Therefore, it was concluded that even for small sensory differences, and in conditions where sensory fatigue could play a greater role (resampling allowed), the tetrad test sill appears like a good alternative to the triangle. Also, the theoretical increase in performance predicted when allowing sample resampling was confirmed.

For the preference study, the same stimuli were evaluated by 208 subjects. Consumer relevance was defined as a significant result between two products in a preference test (assuming no population segmentation). Such significant preferences were found for three out of the four conditions, including the one with the smallest difference for which a significant result had not been found with either the tetrad or triangle. The non-significant preference in the fourth condition was attributed to segmentation in the population.

Therefore, this investigation confirmed further that the tetrad test is a viable alternative to the triangle test, as it exhibits a greater statistical power even in conditions that could potentially affect it negatively. Also, it was shown that a non-significant sensory difference can still result in a significant preference test, outlining the necessity to go beyond the simple use of a 'more powerful' discrimination test when making decisions and to define the actual consumer relevance of an underlying sensory difference.

1. Introduction

Sensory discrimination testing procedures are an integral part of the sensory scientist toolbox. Consumer product manufacturers in the food, beverage and personal care industries are constantly confronted with challenges to modify their products, good examples of which are the current push to reduce salt content or eliminate trans-fat for health related reasons. In this case, like in the majority of situations, the objective is to establish a 'match', i.e., a reformulated product that can be deemed an acceptable alternative to the original product. A scientist can investigate such similarity through two interchangeable routes, one involving power considerations (Ennis, 1993; Ennis & Jesionka, 2011; Schlich, 1993) and the other direct equivalence testing (Bi, 2005; Ennis & Ennis, 2010).

In order to conduct such investigation, a multitude of protocols are available to the sensory scientists: 2-alternative forced choice, triangle, duo–trio, tetrad, two-out-of-five, hexad, octad, dual-pair, same–different, degree of difference, A/Not A, etc. While all these methodologies are used towards the same objective, namely to investigate whether a noticeable difference exists between the products, they vary considerably in their ability to do so in an accurate and efficient manner (Ennis, 1993; Ennis & Jesionka, 2011). A protocol such as the 2-AFC, which generally requires the nature of the reference to be specified in the instructions (e.g., "which of the two samples is harder?"), is statistically more powerful than other

* Corresponding author. Address: The Institute for Perception, 7629 Hull Street Road, Richmond, VA 23235, USA. Tel.: +1 530 753 2231.
E-mail address: benoit.rousseau@ifpress.com (B. Rousseau).

0950-3293/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.foodqual.2013.07.007

more broadly used protocols such as the triangle and duo–trio. Consequently, a researcher is less likely to miss a difference and will have a better estimate of its size by using the 2-AFC. However, since it needs the identification of an attribute, it is usually less straightforward to use than its 'attribute free' counterparts.

The lack of a need to specify an attribute in the triangle and duo–trio methodologies explains why they have been so largely favored in sensory and consumer research. However, their low statistical power regularly results in the miss of sensory differences that can potentially be detected by consumers and lead to the rejection of the reformulated product. In order to address this key weakness, research has been conducted to improve their ability to detect differences: use of replications (Brockhoff & Schlich, 1998; Ennis & Bi, 1998); modification to the experimental protocol (Kim & Lee, 2012; Kim, Lee, & Lee, 2010; Lee & Kim, 2008; Rousseau, Stroh, & O'Mahony, 2002). Also, other protocols that do not require the nature of the difference to be specified in the instructions have been investigated (same–different test (Rousseau, Meyer, & O'Mahony, 1998; Rousseau, Rogeaux, & O'Mahony, 1999; Stillman & Irwin, 1995), 2-AFC-R (Lee, van Hout, & Hautus, 2007; van Hout, Hautus, & Lee, 2011), Torgerson's method of triads (Ennis, Mullen, & Frijters, 1988; Rousseau, 2007; Torgerson, 1958)).

One protocol that seems to be a particularly relevant alternative to the triangle method is the tetrad test (Ennis, Ennis, Yip, & O'Mahony, 1998; Ennis & Jesionka, 2011) with the following instructions: "Here are four samples. Group them into two groups of two based on similarity". While it has been present in the literature for some time, its greater statistical power has only recently been investigated (Ennis & Jesionka, 2011). What was found is that, theoretically, the tetrad methodology is more powerful than the triangle test by a factor of approximately 3 (Ennis, 2012). Since the guessing probability for the tetrad is, like the triangle, $^1/_3$, direct comparisons between the two methodologies can readily be made in terms of performance and the same binomial tables can be used to investigate whether the results of an experiment show a statistically significant difference.

While it exhibits a greater power, the tetrad can potentially suffer from a decrease in performance linked to the addition of a fourth stimulus, compared to the three stimuli comprised in the triangle test. Numerous pieces of research have investigated the effect of memory as well as sensory adaptation and sensitization and have shown that protocols involving fewer stimuli can have a practical advantage over theoretically more powerful procedures (see for instance Dessirier & O'Mahony, 1998; Lau, O'Mahony, & Rousseau, 2004; Rousseau & O'Mahony, 1997). This decrease in sensitivity can be measured in terms of d' values, a standardized measure of sensory difference (Ennis, 1993; Frijters, 1979; Green & Swets, 1966; Macmillan & Creelman, 2005). This effect was for instance investigated using the triangle (three stimuli) and same–different (two stimuli) methodologies, with the triangle's lower d' value being attributed to the greater memory requirements, rather than fatigue or sequence effects (Kim, Jeon, Kim, & O'Mahony, 2006; Lau et al., 2004).

While a decrease in sensitivity, illustrated by a lower d' value, will reduce the tetrad's statistical power, it can still be more powerful than the triangle if this decrease is 'small enough'. Ennis (2012) provides a useful rule that outlines when the tetrad can be a good alternative to the triangle. Ennis showed that the tetrad test is a better alternative to the triangle if its d' value is greater than at least $^2/_3$ of that of the triangle. It is thus valuable to get experimental insights on the relationship between the tetrad and triangle tests' sensitivities, to corroborate the value of the four sample protocol.

Several pieces of research are already available in the literature that compared the triangle and tetrad procedures. Delwiche and O'Mahony (1996) and Garcia, Ennis, and Prinyawiwatkul (2012) indeed found a lower d' value for the tetrad compared to the triangle. However, the decrease of performance was not large enough for the tetrad to lose its advantage. Masuoka, Hatjopoulos, and O'Mahony (1995), working with beer products, did not observe this lower discrimination. However, since two different groups of nine evaluators performed the two protocols, the lack of decrease in performance can possibly be assigned to differences in sensory acuity between the two groups of subjects. Of the above three studies, only that by Garcia et al. (2012) allowed re-tasting and only if all stimuli were tasted in the same order.

The research presented here proposes to expand our current knowledge related to triangle and tetrad testing in three important ways:

(1) Compare the two methodologies specifically. The findings currently reported in the literature mentioned above include the investigation of other methodologies such as the 3-AFC and specified method of tetrads. This in turn could have resulted in the dampening of the sensitivity differences between the protocols due to additional experimental variance, specifically a potential sensitivity advantage of the triangle test might have been concealed because of the additional noise in the overall experiment (sensory fatigue, confusion with different sets of instructions, learning effects upon repeated exposure to the stimuli, etc.).
(2) Use a range of sensory effect sizes for the comparison, especially smaller sizes that are more relevant in situations where a 'match' between products is of interest. The research mentioned above often involved fairly large sensory differences between samples (e.g., $d' \approx 1.3$ in Masuoka et al. (1995), $d' \approx 2.2$ in Delwiche and O'Mahony (1996); $d' \approx 0.8$–1.8 in Garcia et al. (2012)). In order to put these differences into context, in a 2-AFC a d' of 1 corresponds to 76% of tests correct while a d' of 2 corresponds to 92% of tests correct. It is possible that when trying to discriminate between samples with smaller perceivable sensory differences that could be of relevance to the consumer ($d' \approx 0.5$ or 64% of tests correct in a 2-AFC), the addition of a fourth sample in the tetrad protocol will have a greater weakening effect on discrimination due to more 'damaging' adaptation, fatigue or/and memory effects.
(3) Investigate the effect of resampling on overall discrimination when comparing the two protocols. Resampling has been shown to increase discrimination (Caroselli, 2012; Juslin & Olson, 1997; Rousseau & O'Mahony, 2000). However, there is a possibility that when subjects are allowed to retaste and choose to do so, their sensory acuity deteriorates faster in the tetrad condition, thus exacerbating the difference with the triangle procedure.

Building on this first investigation, the second part of this research focuses on a question essential to any successful discrimination testing program: while the tetrad test might be more powerful than the triangle test, can it always successfully detect a difference that consumers would find relevant? Would switching to the tetrad test solve all issues associated with the lack of power often encountered in discrimination testing? Study 2 was to illustrate the fact that simply using a more powerful protocol might not be enough to reach a suitable business decision based on discrimination test results. In order to answer these questions, one must first define 'consumer relevance'. Various metrics are available and here we will consider that the size of a difference is relevant if consumers express a preference for one product over the other. The idea is that if no sensory difference exists between the samples, a consumer population cannot have a preference. As the size of the underlying difference increases, the likelihood of a preference arising increases until it reaches a threshold above which preferences

will be measured.[1] Previous research (Geelhoed, MacRae, & Ennis, 1994; MacRae & Geelhoed, 1992) showed that even though no statistical sensory difference was found when comparing tap and distilled water, consumers could detect a difference when required to identify their preferred stimulus, which was tap water.

To that end, the samples were evaluated under the same conditions (resampling allowed or not), but this time focusing on consumers' preferences for the products. The results were then linked back to those obtained in the triangle and tetrad investigation.

2. Study 1: triangle vs. tetrad

2.1. Materials and methods

2.1.1. Products

The base stimuli were apple and orange juices from Cascadian Farm (Rockport, WA). Within each flavor, the difference was to be created through a specific level of dilution. Preliminary testing using successive dilutions was conducted using members of the lab to estimate approximately the level of dilution for each flavor that would result in the appropriate level of difference between the two samples being compared. Accordingly, a 10% dilution (with distilled water) was chosen for the apple juice, while a 20% dilution was chosen for the orange juice. Samples were dispensed in 10 mL aliquots (primer, no-resampling condition) or 20 mL aliquots (resampling condition) in 2 oz (~60 mL) black plastic cups. Samples were served at room temperature (approx. 16–20 °C), constant within a session.

2.1.2. Subjects

A total of 456 subjects (200 males, 256 females; 15–78 years old, average age 24.4 years old, median 21 years old) participated in this research. They were students, staff, friends and family at the University of California in Davis.

2.1.3. Procedures

Interviews were conducted one on one, to ensure that the instructions were followed precisely. Upon seating at the tasting table across from the interviewer, respondents were asked to rinse their mouth three times with distilled water. They were then presented with four successive sets of samples. Each respondent evaluated two sets in the no-resampling condition and two sets in the resampling condition. Within a resampling condition (allowed or not allowed), the same flavor (apple or orange) was involved. Both a triangle and tetrad tests were performed in each of the two conditions.

For each set, respondents started by taking a diluted sample (the 'primer') to prepare their mouth to the taste of the samples after rinsing with water. They were told that this first sample was not part of the test and that they should not pay attention to its taste. In the no-resampling condition, respondents were required to taste the samples from left to right and evaluate the whole 10 mL of each sample and give their answer upon tasting the last one. In the resampling condition, they were told that they should taste each sample once first (20 mL were served), and then go back if they chose to do so before giving their answer.

For the triangle test, the instruction was to select the sample that was different from the other two. For the tetrad test, respondents were asked to divide the four samples into two groups of two based on similarity. Instructions and responses were given verbally. Testing lasted approximately 10 min on average.

Sample presentation orders, test sequences and resampling conditions were carefully balanced throughout the whole experiment. Half the respondents (228) started with the no-resampling condition, while the other half started with the resampling condition. For each subgroup, half of them (114) started with the orange samples, while the other half started with the apple samples. Each of the six possible presentation orders of both the triangle and tetrad tests was presented 38 times.

2.2. Results

2.2.1. Analyses by condition

The results combined over all subjects are provided in Table 1. d' values and their variances were estimated using IFPrograms™ version 8.10 (The Institute for Perception, USA), which was also used to conduct d' tests between the relevant d' values. This information can also be obtained from tables and instructions available in the literature (Bi, Ennis, & O'Mahony, 1997; Ennis, 1993; Ennis & Jesionka, 2011; Ennis et al., 1998).

For the four conditions, the tetrad resulted in greater numbers of correct responses than the triangle. This is an illustration of its greater statistical power, as predicted by the Thurstonian theory.

For the triangle test in the apple no-resampling condition, the d' value was estimated to be 0, as the proportion of correct answers is slightly lower than chance (0.325). Since no d' variance can be estimated for the triangle test at 0 (it is infinite), no triangle vs. tetrad d' comparison test was possible for that condition. Since the tetrad results are not significantly different from 0 (binomial, $p = 0.14$), it can be concluded that the tetrad and triangle d' values are not significantly different. For the other three conditions, the triangle and tetrad values were not significantly different, even if the triangle test exhibited a slightly higher d' value in each case.

Based on Ennis (2012) research, the tetrad will be a better alternative to the triangle test even if its estimated d' value is lower than that of the triangle as long as the ratio of the tetrad d' over the triangle d' is greater than $^2/_3$. Using Ennis and Ennis (2011) approach to ratio comparisons which uses both the d' values and their associated variances, the ratio can be estimated and compared to $^2/_3$ as a reference point. The confidence levels that the ratio is greater than $^2/_3$ are shown in Table 2.

Therefore, even for relatively small effect sizes (d' of 1 or lower), we would conclude that the tetrad test is a good alternative to the triangle test. For the smallest degree of difference (no-resampling condition for the apple juice), the tetrad even exhibited a performance above chance, while the triangle did not.

2.2.2. Effect of resampling

The effect of resampling can be investigated here as well. It is worth noting that subjects were not required to retaste the samples in the resampling condition. Nevertheless, a vast majority elected to do so. This analysis was conducted two ways:

- The first analysis consists of comparing the d' values obtained from the various conditions, as reported in Table 1. A limitation in this analysis is that the consumers who evaluated the apple juices under the no-resampling condition, for instance, were different from those who evaluated the same products under the resampling condition. Therefore, differences in sensory acuity could confound the effect investigated.
- The second analysis consists of combining the apple and orange data for a given protocol and condition. This raises the sample size for the comparison to 456 and reduces the effect of different sensory acuities in each condition, as all the consumers are included (one could argue that if some consumers have

[1] A relevant level might not result in a significant preference due to population segmentation. However, a significant preference will indicate that the size of the difference is relevant for that group of consumers.

Table 1
Triangle and tetrad results by flavor and resampling condition: number correct, d' and associated variances, binomial probabilities for sample differences and d' value comparisons.

Flavor	Condition	Protocol	# Correct	*N*	d'	σ^2	*Binomial*	d' Comparison
Apple	No resampling	Triangle	74	228	0	N/A	N/A	N/A
		Tetrad	84	228	0.44	0.043	$p = 0.14$	
	Resampling	Triangle	92	228	0.90	0.050	$p = 0.02$	$p = 0.82$
		Tetrad	103	228	0.84	0.017	$p = 0.0001$	
Orange	No resampling	Triangle	96	228	1.02	0.043	$p = 0.003$	$p = 0.65$
		Tetrad	107	228	0.91	0.015	$p < 0.0001$	
	Resampling	Triangle	109	228	1.35	0.032	$p < 0.0001$	$p = 0.32$
		Tetrad	122	228	1.14	0.013	$p < 0.0001$	

Table 2
Triangle and tetrad results by flavor and resampling condition.

Flavor	Condition	Protocol	d'	σ^2	$\frac{d'_{tetrad}}{d'_{triangle}}$	Confidence that ratio is greater than $^2/_3$ (%)
Apple	No resampling	Triangle	0	N/A	N/A	N/A
		Tetrad	0.44	0.043		
	Resampling	Triangle	0.90	0.050	0.93	85
		Tetrad	0.84	0.017		
Orange	No resampling	Triangle	1.02	0.043	0.89	85
		Tetrad	0.91	0.015		
	Resampling	Triangle	1.35	0.032	0.84	90
		Tetrad	1.14	0.013		

different sensory acuity depending on the juice evaluated, the confounding effect would still not be totally eliminated). Results are summarized in Fig. 1.

For all analyses, the resampling d' values are larger than their no-resampling counterparts, confirming predictions and previously observed results. The lack of significance for some of the

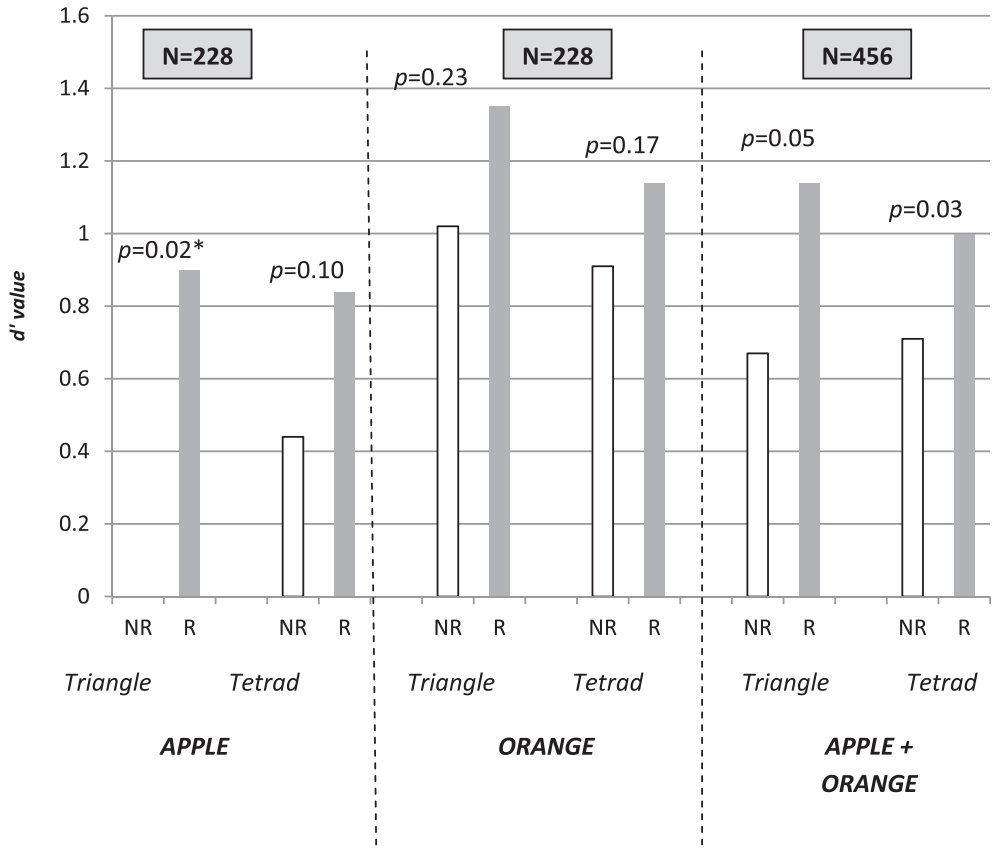

Fig. 1. Comparison of the no resampling and resampling conditions. NR: No resampling allowed condition; R: Resampling allowed condition. *Resampling d' results significantly different from chance level ($d' = 0$).

comparisons can most likely be assigned to a lack of power due to an insufficient sample size.

3. Study 2: product preference investigation

3.1. Materials and methods

3.1.1. Products

The stimuli were the same as those used in Study 1 and the preparation protocol was also identical.

3.1.2. Subjects

Subjects were recruited from the same pools as in Study 1. A total of 208 subjects (94 males, 114 females; 16–59 years old, average age 27.2 years old, median 22 years old) participated in this research. They were students, staff, friends and family at the University of California in Davis.

3.1.3. Procedures

Interviews were again conducted one on one, to ensure that the instructions were followed precisely. Respondents first rinsed their mouth three times with distilled water. Each respondent then performed two successive paired preference tests: one in the resampling condition, the other in the no-resampling condition. Also, one involved the apple juice stimuli, while the other involved the orange juice stimuli.

As in Study 1, respondents first took 10 mL of the primer and were told not to pay attention to its taste. They then tasted the two samples for preference from left to right. In the no-resampling condition, they took the whole 10 mL of each sample and gave their answer upon tasting the second one. In the resampling condition, they were told that they should taste each stimulus once and that they could resample them if they wished to do so (20 mL were served). For each pair, they were asked which of the two samples they preferred, or if they had no preference. Testing lasted approximately 5 min on average.

As in Study 1, sample presentation orders, test sequences and resampling conditions were carefully balanced throughout the whole experiment. Half the respondents (104) started with the no-resampling condition, while the other half started with the resampling condition. For each subgroup, half of them (52) started with the orange samples, while the other half started with the apple samples. Each of the two possible presentation orders (AB, BA) was presented 26 times for each flavor and resampling condition.

3.2. Results

Results are summarized in Table 3. The data were analyzed using the 2-alternative choice model, which makes use of all the data (Braun, Rogeaux, Schneid, O'Mahony, & Rousseau, 2004; Christensen, Lee, & Brockhoff, 2012) and can be used for preference tests involving a 'No preference' option. The 2-AC analyses were again conducted using IFPrograms 8.10.

Other analyses consisting of ignoring, splitting equally or splitting proportionally the no preference answers (see Ennis & Ennis, 2012a; Ennis & Ennis, 2012b) result in the same conclusions as those reached using the 2-AC analysis in terms of significant differences. In general, consumers tended to prefer the concentrated samples for both flavors and both resampling conditions.

Note: while d' values can also be estimated using a preference test, as shown in Table 3, their meaning will be different from d' values obtained from difference tests in the presence of preference segmentation. For instance, even if a large difference exists (e.g., $d' = 2.0$) and is picked up by a difference test, a preference test could result in a smaller d' value being measured and no significant preference. This would be driven by preference segmentation in the population. An extreme situation will be where 50% of the population prefers A, 50% of the population prefers B, i.e., preference $d' = 0$, but products A and B can be clearly sensorily different. Generally, one would expect a preference d' value to be equivalent to or lower than, but not greater than, a d' between the same products obtained in a discrimination test.

4. Discussion

In this research it was hypothesized that while the tetrad test sensitivity did not decrease enough to lose its theoretical power advantage over the triangle test when sensory differences were larger ($d' > 1$), it might not be the case when the underlying difference was less than a d' of 1. This was not confirmed here. The stimuli and testing conditions used in this experiment allowed for the targeting of such smaller d' values in three of the four conditions. In those three conditions, as well as in the fourth where d' values were 1.14 and 1.35, the tetrad test resulted in a higher number of correct answers than the triangle.

For the smallest d', a protocol feature could have played a role in the somewhat counterintuitive trend observed: the size of the estimated d' value variance. For a sample size of 228 and a d' of 0.44, the tetrad d' variance is 0.043. For the same sample size and the same d' value of 0.44, the triangle d' variance would be 0.16. Therefore, even if the triangle testing would generally result in a slightly higher d' value due to more limited experimental variance (as seen in the other three conditions), the imprecision of the d' estimate with its larger variance negates that potential advantage. Another example of this effect is visible in Table 1, where it can be seen that in order for the triangle d' to have approximately the same variance as that of the 0.44 tetrad d' (variance = 0.043), it needs to be more than twice as large (apple juice resampling condition, $d' = 0.90$, variance = 0.05). This illustrates a further benefit of the tetrad over the triangle, namely the higher precision in the estimation of the size of the underlying difference between the products.

Nevertheless, the expected trend is observed in the other three conditions: the tetrad d' value, while not significantly so, is slightly lower than that of the triangle. The analysis based on Ennis (2012) shown in Table 2 confirms that the results show at the 85% confi-

Table 3
Paired preference results by flavor and resampling condition.

Flavor	Condition	Preference response				2-AC analysis		
		Concentrated	No. pref.	Diluted	% Expressed preference* (%)	d'	σ^2	p
Apple	No resampling	57	24	23	71/29	0.63	0.03	<0.001
	Resampling	57	12	35	62/38	0.38	0.03	0.02
Orange	No resampling	54	13	37	59/41	0.30	0.03	0.08
	Resampling	77	3	24	76/24	0.98	0.04	<0.001

* % Preference for concentrated/% preference for diluted.

dence level a decrease in d' value that is less than $\frac{1}{3}$, level at which the tetrad is no longer a relevant alternative to the triangle. Consequently, based on the results reported here, we can confirm the finding from previous research, but this time with smaller sensory differences, and expand the experimental evidence that the tetrad test is a more powerful alternative to the triangle procedure. It is worth noting that using products more fatiguing than fruit juices might yield different outcomes, specifically a greater sensitivity reduction for the tetrad test. This will be the topic of further experimentation.

Regarding the no-resampling vs. resampling conditions, the results confirmed that better discrimination can be achieved by letting subjects resample the stimuli. While the results were not significant with the smaller sample sizes (Fig. 1, $N = 228$), they all show the same trend, namely a higher number of correct answers, and thus d' value, in the resampling condition, even if different groups of consumers are involved in the comparison. When combining the apple and orange juice data, thus involving the same consumers and doubling the sample size, the resampling condition exhibits a significantly higher d' value and thus sensitivity.

The first experiment confirmed that the tetrad test is a more powerful alternative to the triangle test, even for small sensory differences. However, is it powerful enough for all situations? The results from the second experiment showed it is not systematically the case. As mentioned previously, a significant or non-significant result *in itself* in a difference test between two products does not guarantee a suitable prediction of consumers' reaction to the change. For instance, a common use for discrimination tests is the objective of confirming that a reformulation is equivalent to the product it is trying to match. In that case, assuming a 'difference test' rather than an 'equivalence test' approach, it is often assumed that a non-significant difference will signal that the products are indeed equivalent. This is the conclusion that would be reached in Study 1 based on the tetrad test for the apple no-resampling condition. However, it is possible that due to a lack of statistical power in the experiment, a difference that matters to the consumers would be missed.

This is what the research here illustrates. As seen before, based on the tetrad test in the apple juice no-resampling condition, we would conclude that 228 consumers could not detect a significant difference between the samples, and thus possibly confirm that the products are equivalent. However, when looking at the results from the preference test reported in Table 3, we see that a separate group of 104 consumers, recruited from the same population, had a clear preference for the concentrated sample ($p < 0.001$). Therefore, a non-significant difference in the tetrad test with a sample size of 228 can still correspond to a situation where consumers exhibit a strong preference. Assuming that a company only conducted a difference test and found the non-significant result above, it would have released the reformulation that could then have been rejected by consumers who would significantly prefer the original recipe. This is the effect reported by Macrae and Geelhoed (1992). It is worth noting that the tetrad test and the preference test measured the same size of a difference when translated into d' values. The tetrad d' value (0.44) and the preference test d' (0.63) are not significantly different ($p = 0.48$). Finally, using power calculations as described by Ennis and Jesionka (2011), in order for the tetrad test to reach a power of 80% for an alpha of 5% and a δ of 0.54 (average d' value between the tetrad and preference tests), a sample size of 568 would be necessary for the tetrad test. On the other hand, a preference test (modeled as a 2-alternative forced choice) would need a sample size of 78 for the same specifications. The results observed in this research illustrate these predictions. Depending on the target size of the underlying difference, the tetrad test might still require large sample sizes, even if the triangle test's sample size requirements would even be larger (2104 for the above specifications). If the target δ value is greater than 0.54, the required sample size for the tetrad will be lower. However, this shows the importance of investigating and setting the relevant δ value to ensure sufficient power and accuracy in the decision making process.

As for the other experimental preference conditions, they showed similar trends, even if in the orange no-resampling condition the preference was not significant. This can be explained by segmentation in the population, which was confirmed by spontaneous comments from consumers.

To conclude, the difference and preference results reported here show the importance of going beyond relying solely on a test's significance level to make decisions, and to start thinking in terms of the size of the sensory difference relevant to the consumer: the size of the sensory difference below which the products can truly be considered as 'equivalent'. The future of difference testing and the improvements reached when using discrimination protocols rely on establishing the size of a relevant difference for a company's line of products. To that end, the tetrad is a better alternative to the traditional triangle and duo–trio protocols, as it permits greater precision in the estimation of underlying sensory differences between products. However, suitable research is necessary to ensure that the specifications of the protocol used to study product differences, even if that protocol is the tetrad test, are appropriate to deliver the necessary power and warrant that sensory differences relevant to a company's consumers do not go unnoticed.

Acknowledgements

The authors would like to thank Suzanne Pecore of General Mills for facilitating the procurement of the Cascadian Farms products that were used in this research, as well as all the UCD students and staff whose efforts permitted this research to be conducted successfully. We also would like to thank John Ennis for conducting the confidence bound analyses on the ratio of the tetrad and triangle d' values.

References

Bi, J. (2005). Similarity testing in sensory and consumer research. *Food Quality and Preference, 16*, 139–149.

Bi, J., Ennis, D. M., & O'Mahony, M. (1997). How to estimate and use the variance of d' from difference tests. *Journal of Sensory Studies, 12*, 87–104.

Braun, V., Rogeaux, M., Schneid, N., O'Mahony, M., & Rousseau, B. (2004). Corroborating the 2-AFC and 2-AC Thurstonian models using both a model system and sparkling water. *Food Quality and Preference, 15*, 501–507.

Brockhoff, P. B., & Schlich, P. (1998). Handling replications in discrimination tests. *Food Quality and Preference, 9*, 303–312.

Caroselli, A. (2012). Investigation of the effect of allowed and forced within trial retasting on judge performance in the 2-AFC. M.S. Thesis, University of California, Davis.

Christensen, R. H. B., Lee, H.-S., & Brockhoff, P. B. (2012). Estimation of the Thurstonian model for the 2-AC protocol. *Food Quality and Preference, 24*, 119–128.

Delwiche, J., & O'Mahony, M. (1996). Flavour discrimination – An extension of Thurstonian paradoxes to the tetrad method. *Food Quality and Preference, 7*, 1–5.

Dessirier, J.-M., & O'Mahony, M. (1998). Comparison of d' values for the 2-AFC (paired comparison) and 3-AFC discrimination methods: Thurstonian models, sequential sensitivity analysis and power. *Food Quality and Preference, 10*, 1–8.

Ennis, D. M. (1993). The power of sensory discrimination methods. *Journal of Sensory Studies, 8*, 353–370.

Ennis, J. M. (2012). Guiding the switch from triangle testing to tetrad testing. *Journal of Sensory Studies, 27*, 223–231.

Ennis, D. M., & Bi, J. (1998). The beta-binomial model: Accounting for inter-trial variation in replicated difference and preference tests. *Journal of Sensory Studies, 13*, 389–412.

Ennis, J. M., & Ennis, D. M. (2010). Equivalence hypothesis testing. *Food Quality and Preference, 21*, 253–256.

Ennis, J. M., & Ennis, D. M. (2011). Confidence bounds for multiplicative comparisons. *Communications in statistics – theory and methods, 40*, 3049–3054.

Ennis, D. M., & Ennis, J. M. (2012a). Accounting for no difference/preference responses or ties in choice experiments. *Food Quality and Preference, 23*(1), 13–17.

Ennis, J. M., & Ennis, D. M. (2012b). A comparison of three commonly used methods for treating no preference votes. *Journal of Sensory Studies, 27*(2), 123–129.
Ennis, J. M., Ennis, D. M., Yip, D., & O'Mahony, M. (1998). Thurstonian models for variants of the method of tetrads. *British Journal of Mathematical and Statistical Psychology, 51*, 205–215.
Ennis, J. M., & Jesionka, V. (2011). The power of sensory discrimination methods revisited. *Journal of Sensory Studies, 26*, 371–382.
Ennis, D. M., Mullen, K., & Frijters, J. E. (1988). Variants of the method of triads: Unidimensional Thurstonian models. *British Journal of Mathematical and Statistical Psychology, 41*, 25–36.
Frijters, J. E. R. (1979). The paradox of discriminatory non-discriminators resolved. *Chemical Senses and Flavor, 4*, 355–358.
Garcia, K., Ennis, J. M., & Prinyawiwatkul, W. (2012). A large-scale experimental comparison of the tetrad and triangle tests in children. *Journal of Sensory Studies, 27*, 217–222.
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. *Perception & Psychophysics, 55*, 473–477.
Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley & Sons.
Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: a sampling model of confidence in sensory discrimination. *Psychological Review, 104*, 344–366.
Kim, H.-J., Jeon, S. Y., Kim, K.-O., & O'Mahony, M. (2006). Thurstonian models and variance I: Experimental confirmation of cognitive strategies for difference tests and effects of perceptual variance. *Journal of Sensory Studies, 21*, 465–484.
Kim, M. A., & Lee, H. S. (2012). Investigation of operationally more powerful duo–trio test protocols: Effects of different reference schemes. *Food Quality and Preference, 25*, 183–191.
Kim, M. A., Lee, Y. M., & Lee, H. S. (2010). Comparison of d' estimates produced by three versions of a duo–trio test for discriminating tomato juices with varying salt concentrations: The effects of the number and position of the reference stimulus. *Food Quality and Preference, 21*, 504–511.
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 & Psychophysics, 66*, 464–474.
Lee, H.-S., & Kim, K. O. (2008). Difference test sensitivity: Comparison of three versions of the duo–trio method requiring different memory schemes and taste sequences. *Food Quality and Preference, 19*, 97–102.
Lee, H.-S., van Hout, D., & Hautus, M. (2007). Comparison of performance in the A-Not A, 2-AFC, and same–different tests for the flavor discrimination of margarines: the effect of cognitive decision strategies. *Food Quality and Preference, 18*, 920–928.
Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Lawrence Elbaum Associates, Publishers.
MacRae, A. W., & Geelhoed, E. N. (1992). Preference can be more powerful than detection of oddity as a test of discriminability. *Perception & Psychophysics, 51*, 179–181.
Masuoka, S., Hatjopoulos, D., & O'Mahony, M. (1995). Beer bitterness detection: Testing Thurstonian and sequential sensitivity analysis models for triad and tetrad methods. *Journal of Sensory Studies, 10*, 295–306.
Rousseau, B. (2007). Simultaneous estimations of multiple product similarities using a new discrimination protocol. *Journal of Sensory Studies, 22*, 533–549.
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. (1997). Sensory difference tests: Thurstonian and SSA predictions for vanilla flavored yogurts. *Journal of Sensory Studies, 12*, 127–146.
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., Rogeaux, M., & O'Mahony, M. (1999). Mustard discrimination by same–different and triangle tests: Aspects of irritation, memory 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.
Schlich, P. (1993). Risk tables for discrimination tests. *Food Quality and Preference, 4*, 141–151.
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.
Torgerson, W. S. (1958). *Theory and methods of scaling*. New York, NY: Wiley.
van Hout, D., Hautus, M. J., & Lee, H. (2011). Investigation of test performance over repeated sessions using signal detection theory: Comparison of three nonattribute-specified difference tests 2-AFCR, A-NOT A and 2-AFC. *Journal of Sensory Studies, 26*, 311–321.