Trends in Food Science & Technology 66 (2017) 146–152

Contents lists available at ScienceDirect

Trends in Food Science & Technology

journal homepage: http://www.journals.elsevier.com/trends-in-food-science-and-technology

Review

The evolution of paired preference tests from forced choice to the use of 'No Preference' options, from preference frequencies to d' values, from placebo pairs to signal detection

Michael O'Mahony [a], Sukanya Wichchukit [b, *]

[a] Department of Food Science and Technology, University of California, Davis, CA 95616, USA
[b] Department of Food Engineering, Kasetsart University, Kamphaeng Saen Campus, 73140, Thailand

ARTICLE INFO

Article history:
Received 1 November 2016
Received in revised form
6 March 2017
Accepted 23 May 2017
Available online 1 June 2017

Keywords:
Paired preference
Placebo pairs
Signal detection
d'
Signal
Noise

ABSTRACT

Background: Paired preference tests are one of several types of measurement of consumer acceptance. Yet, the test has had issues regarding its statistical analysis. Initially, test designs were 'forced choice', without a 'No Preference' option. Accordingly, the data were limited. Later, putatively identical stimuli were used as controls. Consumers reported preferences for these stimuli. It was deemed not logically possible to have genuine preferences for identical stimuli, therefore these responses were assumed to be responses elicited in the 'no preference' condition. This provided the control condition for statistical comparison. It also allowed responses of 'no preference'. Subsequent experimental designs were refinements of this approach.

Scope and approach: Difficulties with earlier forced choice designs are described and how statistical analysis changed with the introduction of control groups to represent the 'no preference' condition. It describes how the measurements were refined by supplementing frequency measures with d' measures from signal detection theory. The factors causing consumers to report preferences for putatively identical stimuli are discussed. How these have spawned alternative protocols for paired preference tests are described.

Key findings and conclusions: Forced choice preference tests were adopted so that simple binomial statistics could be used. However, they did not allow consumers to express 'no preference'. The introduction of control groups, using putatively identical stimuli, solved this problem. The supplementation of frequency measures with d' measures allowed the use of signal detection protocols, which elicited more accurate measures of preference strength. This is still work in progress; further developments are expected.

1. Introduction

Measurement of consumer acceptance is one of the most important sensory measures for a company to make. There are many types of such measurement, some of which employ rating scales of one type or another (Lim, 2011; Wichchukit & O'Mahony, 2015). However, another type of measurement merely requires consumers to express a preference for one product or another. This is the paired preference test (e.g. Kemp, Hollowood, & Hort, 2009; Lawless & Heymann, 2010; Minoza-Gatchalian & Divino-Brannan, 2009; Moskowitz, Beckley, & Resurreccion, 2006; Resurreccion, 1998). As a casual test, it is a simple tool for preliminary experimentation. Yet, for more formal testing that might be used for decisions regarding whether to continue with a product development program or to launch a product on to the market, there are some issues that must be addressed. One of these issues involves appropriate experimental design with suitable statistical analysis.

Many textbooks tend to describe and illustrate the paired preference test as a forced choice test. By this was meant that consumers when presented with two products, call them 'A' and 'B', are only given two response options. They can state that they prefer 'A' or they prefer 'B'; they may not state that they have no preference. The reason for this was to allow the test to be analyzed using simple binomial statistics. Yet, this causes problems. It is quite reasonable to expect consumers to express no preference when

* Corresponding author.
E-mail address: fengskw@ku.ac.th (S. Wichchukit).

http://dx.doi.org/10.1016/j.tifs.2017.05.017
0924-2244/

given two food products. Also, should a sample of consumers be equally split in their preferences (50% prefer 'A' and 50% prefer 'B') then it is difficult to interpret the results. Either none of the consumers have a preference and are responding randomly or half the sample prefer 'A' and half prefer 'B'. Without a 'No Preference' option there is no way of knowing which is true, unless it can be determined during subsequent questioning of consumers regarding the reasons for their preference. However, this is not always done. There are examples in the literature where a split result has been wrongly interpreted as consumers having no preference. Therefore, it is safer to allow consumers a 'no preference' option.

2. Excluding no preference

Yet, the textbook authors generally found difficulty in choosing a suitable statistical analysis for a test with three response options. Instead, they described ways of fitting trinomial data into a binomial test by disposing of any 'No Preference' responses (Kemp et al., 2009; Minoza-Gatchalian & Divino-Brannan, 2009; Moskowitz et al., 2006; Resurreccion, 1998). Three ways were usually suggested for dealing with unwanted 'No Preference' responses. Firstly, they could simply be ignored. Secondly, they could be split 50:50 between the two preference options. Thirdly, they could be split proportionately according to the ratio of responses in each preference option. Ennis and Ennis (2012b) modeled these three ways of disposing of 'No Preference' responses in terms of power and reported that the third method of proportionate splitting was more likely to elicit Type I errors. Faulk, Henrikson, and Morrison (1975) added a fourth way; they tossed a coin and whether it landed 'heads' or 'tails' would determine to which set of preference responses the 'No Preference' responses would be added. Yet, all these methods are essentially dishonest. They alter data by attributing preferences to consumers who have no preferences, simply for the convenience of using binomial statistics. Altering data risks drawing the wrong conclusions, see later.

Providing further support for the forced choice protocol, Kemp et al. (2009) reasoned that with the forced choice procedure, no data was lost so that the analysis was statistically more powerful. The results of Odesky (1967) based on consumer surveys also gave support. These indicated that consumers who had expressed no preference would, if forced to choose, distribute their preferences according to those consumers who had used the forced choice test. However, Angulo and O'Mahony (2005) repeated and extended Odesky's experiment and found his assertion to be incorrect. Perhaps the most forceful support for the forced choice protocol, came from Meilgaard, Civille, and Carr (1991) who stated that only the forced choice technique was amenable to formal statistical analysis.

Ignoring or disposing of no preference options would not do much harm if the number of 'No Preferences' were suitably small. However, this is not always the case (Angulo & O'Mahony, 2005; Villegas-Ruiz, Angulo, & O'Mahony, 2008). It is also worth noting that a consistent position in the legal profession is that unless a 'no preference' option is explicitly provided, one cannot be sure how results have been affected by those who were forced to report a preference when they did not have one (Ennis & Ennis, 2012a). Stone and Sidel (2004), when describing the forced choice version of the test, added options that were similar to the 'No Preference' option; they used 'like both equally' and 'dislike both equally'. Yet, in their description of the appropriate statistical analysis, these options were not used; they used the regular binomial statistical analysis. Lawless and Heymann (2010) suggested using a Thurstonian analysis (Lee & O'Mahony, 2004; O'Mahony, Masuoka, & Ishii, 1994) based on the 2-AC test (Braun, Rogeaux, Schneid, O'Mahony, & Rousseau, 2004). This is discussed later.

3. Preferences with putatively identical products

The beginnings of a solution for a suitable statistical analysis had actually been available for some years in an internal report to the tobacco industry. Ennis and Collins (1980, p. 50) mailed various pairs of cigarettes (call them 'A' and 'B') to approximately 2000 consumers for comparison on a wide variety of attributes like: better flavor, slower burning, easier draw etc. Finally, they were asked to report which cigarette in the pair they preferred. In response, 40% reported preference for cigarette 'A', 40% reported preference for cigarette 'B' and 20% reported 'No Preference'. Yet, the pairs of cigarettes had been taken from the same manufacturing run; they were essentially the same cigarette. Accordingly, it was concluded that any preferences expressed for one or other of the cigarettes would have had to be due to factors other than their sensory characteristics.

From then on, the advances in paired preference testing were mainly the province of academia. Marchisano et al. (2003) noted that the response frequencies recorded in their research did not resemble the 40-20-40 results of Ennis and Collins (1980). They found that the proportion of 'no preferences' in the putatively identical pair varied depending on the product being tested, the wording chosen for questioning the consumer, the number and types of no preference responses given on the response sheet, and the culture of the consumer. They found that Korean consumers were more prone to reporting preferences for the 'identical' pair than Americans. In later research (Kim, Lee, O'Mahony, & Kim, 2008; Sung, Lee, O'Mahony, & Kim, 2011), this was confirmed, although the effect was not as strong. For reasons as yet unresolved experimentally, various authors (e.g Alfaro-Rodriguez, O'Mahony, & Angulo, 2005; Alfaro-Rodriguez, Angulo, & O'Mahony, 2008, 2007; Chapman, Grace-Martin, & Lawless, 2006; Chapman & Lawless, 2005; Kim et al., 2008; Sung et al., 2011) reported frequencies that were different from those of Ennis and Collins (1980) depending on the experimental conditions and the consumers tested. The percentage of 'No Preference' responses in the 'identical' pair tended to be higher than 20% ranging as high as 35%. Yet, in all cases, the majority of judges indicated preferences rather than no preference.

The question arose as to what were the factors that elicited reports of preferences with a putatively identical pairs. Marchisano et al. (2003) interviewed consumers who had made such reports. Their replies were varied but they all indicated a strong primary preconception that logically with a preference test, the stimuli would have to be different. A secondary preconception was that consumers were expected to report a preference. The research of Chapman et al. (2006) illustrated these strong preconceptions. They described how panelists were given a pair of identical color swatches (used as examples for paint colors) in a clear plastic bag with the manufacturers color number and color name visible, to indicate that they were the same. Some reported that they could not believe that they could be given identical swatches and then be asked which one they preferred. One panelist had such a strong preconception that she tried to smell the color swatches in her attempts to find a difference. Ennis and Collins (1980) commented that men tended to be more 'courageous' than women in going against their preconceptions and reporting no preference for the 'identical' pair. There was an exception, however. Women who smoked Virginia Slims, a cigarette aimed at young professional women, with marketing themes including independence and liberation, tended to be as 'courageous' as the men, by being less prone to reporting preferences for the 'identical' pair.

From the various explanations based on the consumers' strong preconceptions, three hypotheses emerged to explain why

consumers reported preferences with 'identical' pairs of stimuli: the 'sensory noise' hypothesis, the 'invention' hypothesis and the 'attention' hypothesis. The 'sensory noise' hypothesis assumed that consumers notice a momentary variability in sensory input due to 'noise' in the sensory system, in the same way that they might do for a difference test. This is described by Thurstonian modeling (Lee & O'Mahony, 2004; O'Mahony et al., 1994). For taste, there are several sources of noise. There is neural noise which is caused by the spontaneous firing of additional nerves not concerned with transmitting the signal in question. There are distortions due to the effects of forgetting. Added to these are disturbances in the mouth like adaptation to residual stimuli from prior tastings and effects of stimulus dilution by saliva (Kim, Jeon, Kim, & O'Mahony, 2006; O'Mahony & Goldstein, 1987; O'Mahony & Odbert, 1985). All of these provide a plentiful source of sensory noise which is increased with stronger stimuli and longer intervals between tasting the stimuli.

The 'invention' hypothesis assumes that the consumers do not perceive any differences between the products in the test. However, they are so sure that there must be a difference that they invent one. For example, Marchisano et al. (2003) noted, but did not report in their paper, a consumer, who when presented with pairs of 'identical' chips, persuaded herself that one had to be a low-fat version and gave it her preference. The third hypothesis, the 'attention' hypothesis, also assumes that consumers do not perceive any differences between the products. It utilizes the idea of Kahneman's (2011) 'fast thinking'. It was hinted at by one of the consumers that Marchisano et al. (2003) interviewed. He said that the preference task felt like being asked by a friend, out of politeness, to have first choice of one of an identical pair of coffees that he had bought for them. In the words of Ullmann-Margalit and Morganbesser (1977) he was 'picking' rather than 'choosing', the latter being a selection based on preference, the former being a selection not based on preference. In terms of Kahneman's model, consumers would be making a selection with very little consideration to what was being selected. It has been described as working on 'auto-pilot'. This hypothesis was predicted by a systems analysis of cortical function reviewed in more detail by Xia, Zhong, and O'Mahony (2016a). Such choices made without proper attention, are relegated for convenience to cognitive subroutines rather than central processing. To distinguish between these three hypotheses experimentally will be a challenge.

Along with these hypotheses, Wichchukit and O'Mahony (2011) sounded a warning note. Borrowing a concept from psychology, they defined a preference measured in a formal preference test as a 'test preference' and a preference observed away from the testing situation, in the different circumstances of everyday life, as an 'operational preference'. Accordingly, test preferences are intended to aid in the prediction of operational preferences. However, any preferences reported with an 'identical' pair, would not be sufficiently strong to elicit an operational preference. This was demonstrated experimentally (Prescott, Leslie, Kunst, & Kim, 2005; Xia, Zhong, & O'Mahony, 2015). Arguing by *reductio ad absurdum*, it would lead to people rejecting the second bite of a food or the second sip of a beverage, because a barely perceptible difference might have been perceived and taken as evidence that they were tasting a completely different product. In Thurstonian terms, Wichchukit and O'Mahony's point can be re-stated as the τ-criterion required for preference, needing to be larger than that required for discrimination. To explain this, the τ-criterion is defined as the degree of difference in flavor required for a consumer to feel, in a difference testing context, that a particular very small difference in sensation is sufficient to be reported as real. Yet, in a preference testing context, the τ-criterion is defined as the degree of difference in flavor required for consumer to feel that the two stimuli are sufficiently dissimilar to be regarded as different products, so that preferences might be elicited. This requires a larger difference in sensation. In the jargon of psychology, the τ-criterion is referred to as the measure of 'bias' that the consumer sets for making such a judgement.

Finally, with a 'test preference', lack of a preference is indicated by the consumer choosing the 'No Preference' option. On the other hand, an operational 'No Preference' is indicated by the consumers changing their real life preferences. Köster (Köster, 2003; Köster, Couronne, Léon, Lévy, & Marcelino, 2002) pointed out that consumers change their operational preferences from time to time, making prediction problematic. Consequently, the chances of a 'test preference' failing to predict an 'operational preference' is not so unlikely. Consumers can change their mind from time to time. The research question becomes one of for how long a particular preference measurement protocol can successfully predict an operational preference.

4. Including no preference: placebo pairs, identicality norm

As far as statistical analysis was concerned the main impact of Ennis and Collins's (1980) research was to provide a method of analysis for preference tests with a 'No Preference' option. It was hypothesized that it was not logically possible to have a 'real life' operational preference when the stimuli were identical. Accordingly, the response frequencies for a putatively identical pair were hypothesized to be the response frequencies for the 'No Preference' situation. This gave the statisticians a method for significance testing. Using a chi-squared analysis, the response frequencies for the target (different) pair of products could be compared to the response frequencies derived from the putatively identical pair. A significant difference would indicate that the response frequencies in the target pair were significantly different from the situation where the consumers had no preference. The putatively identical pair came to be called the 'placebo pair' (Alfaro-Rodriguez, Angulo, & O'Mahony, 2007, 2008) for obvious reasons. The preference frequencies derived from the placebo pair were called the 'identicality norm' (Ennis & Ennis, 2012a, b). The 'identicality norm' could be the preference frequencies reported by a sample of consumers or it could be adjusted to represent the frequencies that would be recorded from the population of consumers (Xia, Zhong, & O'Mahony, 2016b).

Accordingly, researchers wishing to use the 'No Preference' option, presented two pairs of stimuli, both a target and a placebo pair. Significance was tested by comparing the response frequencies of the target pair with the placebo pair's identicality norm (e.g. Alfaro-Rodriguez et al., 2008, 2007; Alvarez-Coureaux, Aguilar, O'Mahony, & Angulo, 2010; Marchisano et al., 2003; Sung et al., 2011; Xia et al., 2016b).

For further sophistication, instead of comparing the response frequencies for the target pair and placebo pair, a fundamental measure, d', was adopted. This had been borrowed from signal detection theory, used in communications engineering, (Wichchukit & O'Mahony, 2010a). It was a measure of discrimination between two stimuli. It was a measure of the distance between the sensory distributions associated with two stimuli, in units of standard deviations. In essence, it was a signal-to-noise ratio. In this context, it could be understood as a single unitary measure representing an overall preference tendency of the consumers towards one item or the other (Angulo & O'Mahony, 2005). This would not be always easy to discern from preference frequencies, if there were a substantial number of 'no preference' choices. Also, being a fundamental measure, d' values allowed comparisons between studies using different experimental protocols, for example with and without the 'No Preference' option (Alfaro-Rodriguez

et al., 2005; Angulo & O'Mahony, 2005).

For forced choice preference tests, d' was obtained from the preference frequencies, using tables (Ennis, 1993). For tests with a 'No Preference' option, the computation for the 2-AC difference test was used (Alfaro-Rodriguez et al., 2007; Braun et al., 2004; Lawless & Heymann, 2010; Sung et al., 2011). Accordingly, along with the response frequencies elicited by the various response options, d' values derived from these frequencies were also used for comparisons between target pairs and placebo pairs. The statistics associated with these computations continued to be refined (Christensen, Ennis, Ennis, & Brockhoff, 2014; Christensen, Lee, & Brockhoff, 2012; Ennis & Ennis, 2012a,b; Jesionka, Rousseau, & Ennis, 2014).

5. Borrowing from communications engineering: a signal detection analysis

Using the logic given above, comparisons were made between the response frequencies and d' values for the target pairs and the response frequencies and d' values for the placebo pairs, (the identicality norm). Should there be a significant difference between these two measures it would indicate that the preference responses reported in the target pair were significantly different from the situation where none of the consumers had any preference. Yet, there was a problem. The consumers responding to the target pair were the same consumers as those who had assessed the placebo pair, some of whom had responded with preferences. The question then arises as to what proportion of consumers, when assessing the target pair, were basing their judgments on the sensory input from the products being assessed and what proportion were merely reacting to factors other than their sensory characteristics, as they had been when assessing the placebo pair. This information is necessary if an accurate measure is to be obtained for the strength of preferences observed in the target pair.

One way of approaching this problem was to continue to use the logic of signal detection theory and use the statistical analysis developed by Bi, Lee, and O'Mahony (2015). This analysis developed a Thurstonian model for preference tests with a 'no preference' option, that used both a target and a placebo pair. It yielded a measure of the degree of preference, using d' as well as a measure of the τ-criterion which can be considered as a measure of the bias used in the preference decision (see above). Consider consumers assessing two products in the target pair and reporting a preference represented by a d' value. Consider also that consumers were presented with one or other of the placebo pairs. If the consumers reported a preference for the placebo pair, it would be a reaction to factors other than the sensory characteristics of the stimuli presented in the target pair. As such, it can be considered as a measure of the 'noise' in the system, again represented by a d' value.

Taking this approach, it was now possible to get around the problem of some consumers reacting to noise in the system, when assessing the target pair, rather than the sensory input from the two products being assessed. If the report of a preference in the placebo pair is a measure of the noise in the system, the report of a preference in the target pair can be considered as a measure of 'signal + noise'. The difference between the two, again represented by a d' value, would therefore be a measure of the signal strength of the preference per se, in other words the strength of the preference reported in the target pair, independent of the noise.

Building on this approach Zhang, Halim, Wichchukit, O'Mahony, and Hautus (2016) took the argument a step further, this time using a different, yet typical signal detection computation. They developed a method for measuring the signal strengths for each of the two products in the target pair separately, independent of the noise. Consider a consumer assessing two products, 'A' and 'B', in the target pair and was able to respond with preference for either 'A' or 'B' or having no preference. Consider also that the consumer was presented with a placebo pair comprising either two 'A' stimuli (AA) or two 'B' stimuli (BB). If the consumer reported that 'A' was preferred in the target pair, then the response to the placebo pair 'AA', would be a reaction to factors other than the sensory characteristics of 'A', when 'A' was being selected. As such, it can be considered as a measure of the noise in the system for stimulus 'A'. Again, if the choice of 'A' in the placebo pair is a measure of the noise in the system, the choice of stimulus 'A' in the target pair was considered as a measure of 'signal + noise'. In the same way, the choice of stimulus 'B' in the target pair can be considered as 'signal + noise', while the choice of 'B' in the placebo pair 'BB' was considered as a measure of the noise in the sensory system for that stimulus. With access to the sensory distributions associated with the choice of 'A' in the target pair (signal + noise) and the choice of 'A' in the 'AA' placebo pair (noise), the difference between the two, in terms of d' would be a direct measure of the strength of the preference for 'A' (signal) independent of the noise. The same would be true for stimulus 'B'.

Consider an illustrative example. As is usual with signal detection measures, a preference test protocol using graded responses is required. Instead of choosing between the three response options: 'prefer A', 'prefer B' and 'No Preference', consumers use a graded response protocol, as follows:

I would only choose 'A' and never choose 'B'
I would only choose 'A' but I couldn't promise I would never choose 'B'
I would choose 'A' much more than 'B'
I would tend to choose 'A' a little more than 'B'
I would tend to choose each one roughly the same amount of time

The remaining four preference options are as above, except that 'B' was chosen over 'A'. The consumers find these options easy to deal with; choosing from nine options is no more demanding than the 9-point hedonic scale. Note that in this study preferences are expressed in terms of choosing. This, as well as purchase intent, are more actionable measures of preference. From the response frequencies for the choices of A and B in the target pair and the two placebo pairs, four distributions are constructed. With suitable statistics, theoretical distributions representing the populations from which the sample distributions are derived and fitted to the sample distributions. In this case, let the best fits be normal and have the same variance, so that differences between the means of the distributions, measured in standard deviations, turn out to be measures of d'. By determining d' values representing the difference between the means of the distribution for choosing A in the target pair (signal + noise) and in the AA placebo pair (noise), the real signal strength for choosing A in the target pair can be determined. The same procedure is applied to stimulus B. The four distributions so derived are illustrated in Fig. 1.

For the computation in their experiments, Zhang et al. (2016) used free software available on the website http://hautus.org (SDT Assistant, Version1.0 retrieved from http://hautus.org) for their analysis and they also give a description of how to enter the data. For confirmation, they also used the scaling program in the Institute for Perception's package of programs (mail @ifpress.com, Institute for Perception, Richmond VA) which performs the same analysis.

Returning to Fig. 1, the vertical lines are preference boundaries, which represent the divisions between the verbal response categories shown above. The dashed lines illustrate the distributions for the two placebo pairs. The distribution for placebo 'AA' is on the far

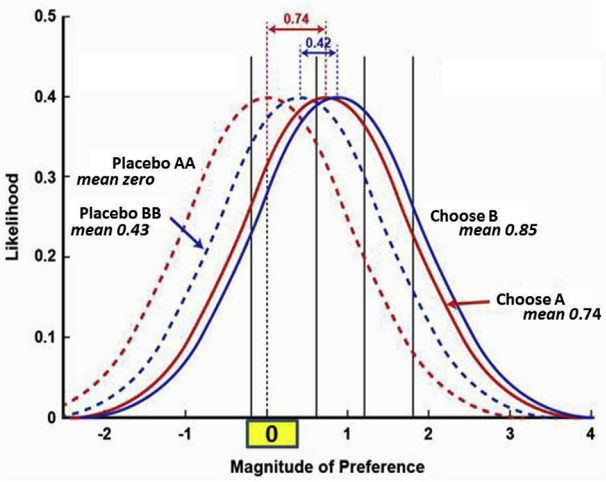

Fig. 1. Hedonic distributions for the two placebo pairs, 'AA' and 'BB' are represented as dashed lines. Hedonic distributions for choosing 'A' or 'B' in the target pair are represented by continuous distributions. The mean of the 'AA' distribution is chosen as the zero for the Magnitude of Preference axis. Mean values for the distributions are shown in the figure.

left. For convenience, the mean of this distribution was chosen as the zero on the Magnitude of Preference axis. As shown on Fig. 1, the mean of the BB placebo distribution is greater (0.43). The two solid line distributions represent the responses of those consumers who chose either 'A' or 'B' in the target pair. The means of these distributions are also indicated in Fig. 1.

The traditional approach for comparing whether consumers preferred to choose 'A' or 'B' with the target pair, would be to determine the d' value between the fitted distributions representing those who chose 'A' and those who chose 'B'. The means of these two distributions on the Magnitude of Preference axis are 0.85 for choosing 'B' and 0.74 for choosing 'A', resulting in a very small d' value of 0.11. Using tables (Bi, Ennis, & O'Mahony, 1997) this is seen not to be significant. Thus, the conclusion would be that there was no significant difference between the tendency to choose 'A' and to choose 'B'.

However, this computation ignores the effects of noise. The distributions for choosing 'A' and 'B' were not signal distributions, they were signal + noise distributions. Accordingly, the appropriate d′ values correspond to the differences between the distributions for choosing 'A' and choosing 'B' and their appropriate placebo pair distributions. As illustrated in Fig. 1, the d' value between choosing 'A' and its placebo is 0.74 (0.74–0); this is significant. The d' value between choosing B and its placebo is 0.42 (0.85–0.43) which is not significant. From this we deduce that the consumer's tendency to choose 'B' is no greater than the noise level. Therefore the new conclusion is that there is a significant tendency for the consumers to choose 'A' but no significant tendency to choose 'B'. The conclusions, when noise is taken into consideration, can be different. Ignoring the effects of noise could possibly lead to erroneous conclusions.

In their actual experiment, Zhang et al. (2016) compared preferences between barbecue flavored chips (B) and sour cream and onion flavored chips (S) using 247 consumers. They fitted distributions to their data as described above. They also fitted distributions with different variances but found them to have no advantage over a fit using the same variance. Once again, using fits of the same variance meant that the differences between the means of the distributions corresponded to d'. The means were determined for the placebo distributions (SS = 0; BB = 0.288) and choice distributions (Choose S = 0.933; Choose B = 1.042). Using the traditional analysis, the two choice distributions were compared giving a d' value of 0.109 (1.042–0.933). This is not significant ($p = 0.568$). The conclusion was that the sample of consumers showed a nonsignificant tendency to choose B over S. The response frequencies indicated that this was the result of two opposing segmented groups. However, this comparison ignored the effects of noise. By subtracting the means of the placebo pairs, the pure signal strengths of the choice values were determined (S = 1.042; B = 0.645). The d' value for the choice of S over can B became larger (1.60), with a lower significance level $p = 0.098$. The conclusion was, once again, that there was a nonsignificant tendency to choose B over S. However, in this case, the preference of S over B was stronger and would even be regarded as significant at the 10% level. Accordingly, the difference in the size of the two opposing segments was seen to be greater.

6. Conclusions

As stated at the beginning of this review, the paired preference test is often used as a casual test for the collection of preliminary data, for example, to obtain some feedback during product development. In this case, the test will be quick and easy and the status of the data obtained treated appropriately. However, even with a 'quick and easy' test, it is important to remember that there will be a strong preconception to report a preference, even when no operational preference exists.

On the other hand, a paired preference test may also be used in the gathering of information for important decisions regarding the continuation of a product development project or the choice of products to release on the market. In this case, it will be important to understand how the area of paired preference testing has advanced, so that a more advanced experimental protocol can be chosen. Accordingly, placebo pairs may be included in the testing protocol as well as the use of a more sophisticated analysis. The greater the precision required by the test, the more is the time required and the more are the controls required. Yet, preference tests do not take long; they are one of the shortest sensory tests. They last only a matter of minutes, even with added placebo tests and various controls. The time required for a consumer to perform the experiment described for Zhang et al. (2016) only ranged 5–10 min and this was one of the longer types of paired preference test. Sensory professional should have sufficient knowledge to be able to choose a test method that is suitable for the products at hand and the importance of the data obtained, although with preference tests the difference in testing time between the casual test and the more sophisticated test is relatively small.

It may be argued that perhaps the most profound change in the development of paired preference tests was triggered by the discovery that consumers would report a high proportion of preferences for putatively identical stimuli (Ennis & Collins, 1980). This introduced placebo pairs and the ability to use a statistical analysis for tests that included the 'no preference' option. Almost as profound was the parallel development of Thurstonian modeling (Ennis, 2016, 1993; Ennis & Jesionka, 2011; Ennis, Rousseau, & Ennis, 2016) which itself is a development of Signal Detection Theory from communications engineering (Green & Swets, 1966; Wichchukit & O'Mahony, 2011). This provided a great degree of sophistication to the statistical analyses with the introduction of d'.

Yet, most of the recent development of this field has been within the confines of academia. It will take time for some of these developments to become routine within the food industry. This is because the latest methodological techniques developed in academia are not always transmitted to the industry in a simple and easy to understand form. In fact, they are often not even taught in

university courses. Also, it is essential that analyses, using concepts like 'signal' and 'noise', are available to the industry in a user-friendly form, so that even sensory professionals without an advanced education in statistics, can use them easily. Only then, will the latest developments in the methodology of paired preference testing be used routinely by the industry.

There is, however, a topic that despite its importance, has been neglected as far as paired preference tests are concerned. It is anticipated that this might be the next development in the evolution of preference tests. It is certainly overdue and is the problem of validity. The important thing to know about a preference test is how well it predicts real life behavior. In the words of Wichchukit and O'Mahony (2011), test preferences are intended to aid in the prediction of operational preferences. Yet, surprisingly, despite its importance, research into validity is an area that has been generally neglected. If predictive tests are being used, it is important to have some idea about how well they predict. Köster is of the opinion that "first hedonic impressions are poor predictors of final liking and choice" (Köster et al., 2002) and "first impression of a new product may have little predictive value for its later success" (Kremer, Shimojo, Holthuysen, Köster and Mojet. 2013). Yet, even if tests are bad predictors it is as well to know how bad they are.

One approach has been to repeat tests. Baker, Amerine Roessler and Filipello (1960) repeated preference tests (forced choice) for raisins on 10 consecutive days and noted that only 31% of consumers were consistent over all 10 tests. Wilke, Cochrane, and Chambers (2006) repeated forced choice preference tests for raisin bran four times during a half hour session. They repeated this for colas. High numbers of consumers change their preference response at least once (50% for bran; 71% for cola). Chapman and Lawless (2005) gave paired preference tests with a single repetition for milk samples and cottage cheeses using the 'no preference' option. They found substantial changes in the consumers' responses (43% for milk; 45% for cheese). Yet, the authors of the latter two papers expressed doubts about how well the consumers could discriminate the products. Certainly this approach is a good idea but is only valid if it has been shown that the consumers could easily discriminate between the products being used in preference tests.

Repetition is used to increase the power of difference tests and appropriate statistics are available. (e.g. Bi & Ennis, 1999a,b; Bi, Templeton-Janik, Ennis, & Ennis, 2000). Yet, if the products are easily discriminable, it may be hypothesized that repetition during the same experimental session would not give consumers time to change their operational preferences. Repetition would only seem sensible over longer periods of time to allow an adequate sampling of real life behavior.

A different approach to testing validity was the 'take-away' test. A preference test was given to consumers after which they were allowed to take away a sample of one or of both products for their use. It was important that the experimenter did not bias a consumer's choice by observing the consumers making their choice. What the consumers took away were compared with the results of the preference test to see the level of agreement. No more than 60% of consumers took away a product or products that corresponded to their response option chosen in the preference test. (Calderón, Angulo, O'Mahony, & Wichchukit, 2015; Weiss, O'Mahony & Wichukit, 2010; Wichchukit & O'Mahony, 2010b, 2011). The authors were careful to state that the 'take-away' test was not an adequate test of validity but it was the small step in the right direction.

Currently, the authors are conducting more comprehensive tests of validity. For this, measures of preference are replaced by measures of likelihood to choose or to buy, because they are more actionable. In a current project, a battery of different types of preference test (choosing) are given to consumers, who then return every week or two and to be offered the same products for them to choose for themselves. Any changes in choice preference are monitored over a three month period, to determine whether they are in accord with predictions of the original tests. This research is not yet completed but the authors hope that it might encourage others to do similar experiments.

So the development of preference tests has not ceased; it still continues. As stated at the beginning, this report is a review of work in progress.

References

Alfaro-Rodriguez, H., Angulo, O., & O'Mahony, M. (2007). Be your own placebo: A double paired preference test approach for establishing expected frequencies. *Food Quality and Preference, 18*, 353–361.

Alfaro-Rodriguez, H., Angulo, O., & O'Mahony, M. (2008). Paired preference tests: '50:50' and 'alternating' no preferences. *Journal of Sensory Studies, 23*, 765–779.

Alfaro-Rodriguez, H., O'Mahony, M., & Angulo, O. (2005). Paired preference tests: d′ values from Mexican consumers with various response options. *Journal of Sensory Studies, 20*, 275–281.

Alvarez-Coureaux, Y., Aguilar, P., O'Mahony, M., & Angulo, O. (2010). Assessment of preference with controls for response bias operating in the test situation: A practical example of using omega-3 enriched wholegrain breads with Ecuadorian consumers. *Journal of Sensory Studies, 25*, 659–671.

Angulo, O., & O'Mahony, M. (2005). The paired preference test and the 'no preference' option: Was Odesky correct? *Food Quality and Preference, 16*, 425–434.

Baker, G. A., Amerine, M. A., Roessler, E. B., & Filipello, F. (1960). The nonspecificity of differences in taste testing for preference. *Food Research, 25*, 810–816.

Be, J., & Ennis, D. M. (1999b). Beta-binomial tables for replicated difference and preference tests. *Journal of Sensory Studies, 14*, 347–368.

Bi, J., & Ennis, D. M. (1999a). The power of sensory discrimination methods used in replicated difference and preference tests. *Journal of Sensory Studies, 14*, 289–302.

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.

Bi, J., Lee, H.-S., & O'Mahony, M. (2015). A Thurstonian model and statistical inference for the 2-alternative choice test with both test pairs and placebo pairs. *Journal of Sensory Studies, 30*, 10–20.

Bi, J., Templeton-Janik, L., Ennis, J. M., & Ennis, D. M. (2000). Replicated difference and preference tests: How to account for inter-trial variation. *Food Quality and Preference, 11*, 269–273.

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.

Calderón, E., Angulo, O., O'Mahony, M., & Wichchukit, S. (2015). "Liking" and "Take Away" preferences for Mexican consumers: Cross-cultural comparison with Thais for psychological style. *Journal of Sensory Studies, 30*, 77–84.

Chapman, K. W., Grace-Martin, K., & Lawless, H. T. (2006). Expectations and stability of preference choice. *Journal of Sensory Studies, 21*, 441–455.

Chapman, K. W., & Lawless, H. T. (2005). Sources of error and the non-preference option in dairy product testing. *Journal of Sensory Studies, 20*, 454–468.

Christensen, R. B. H., Ennis, J. M., Ennis, D. M., & Brockhoff, P. B. (2014). Paired data with a no-preference option - statistical tests for comparison with placebo data. *Food Quality and Preference, 32*, 48–55.

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

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

Ennis, D. M. (2016). *Thurstonian models: Categorical decision making in the presence of noise*. Richmond VA: The Institute for Perception.

Ennis, D. M., & Collins, J. (1980). *The distinction between discrimination and splitting in paired testing*. Report # 80-233. Richmond, Virginia: Philip Morris Research Center.

Ennis, D. M., & Ennis, J. M. (2012a). Accounting for no difference/preference responses or ties in choice experiments. *Food Quality and Preference, 23*, 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*, 123–129.

Ennis, J. M., & Jesionka, V. (2011). The power of sensory discrimination methods revisited. *Journal of Sensory Studies, 26*, 371–382.

Ennis, D. M., Rousseau, B., & Ennis, J. M. (2016). *Tools and applications of sensory and consumer science*. Richmond VA: The Institute for Perception.

Faulk, S. N., Henrikson, R. L., & Morrison, R. D. (1975). Effect of boning beef carcasses prior to cheating on meat tenderness. *Journal of Food Science, 40*, 1075–1079.

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

Jesionka, V., Rousseau, B., & Ennis, J. M. (2014). Transitioning from proportion of discriminators to a more meaningful measure of sensory difference. *Food Quality and Preference, 32*, 77–82.

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus & Giroux.

Kemp, S. E., Hollowood, T., & Hort, J. (2009). *Sensory evaluation a practical handbook*.

Oxford, UK: Wiley-Blackwell.
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, H. S., Lee, H.-S., O'Mahony, M., & Kim, K.-O. (2008). Paired preference tests using placebo pairs and different response options for chips, orange juices and cookies. *Journal of Sensory Studies, 23*, 417–438.
Köster, E. P. (2003). The psychology of food choice: Some often encountered fallacies. *Food Quality and Preference, 14*, 359–373.
Köster, E. P., Couronne, T., Léon, F., Lévy, C., & Marcelino, A. S. (2002). Repeatability in hedonic sensory measurement: A conceptual exploration. *Food Quality and Preference, 14*, 165–176.
Kremer, S., Shimojo, R., Holthuysen, N., Köster, E. P., & Mojet, J. (2013). Consumer acceptance of salt-reduced "soy sauce" bread over repeated in home consumption. *Food Quality and Preference, 28*, 484–491.
Lawless, H. T., & Heymann, H. (2010). *Sensory evaluation of food* (2nd ed.). New York: Springer.
Lee, H.-S., & O'Mahony, M. (2004). Sensory difference testing: Thurstonian models. *Food Science and Biotechnology, 13*, 841–847.
Lim, J. (2011). Hedonic scaling: A review of methods and theory. *Food Quality and Preference, 22*, 733–747.
Marchisano, C., Lim, J., Cho, H. S., Suh, D. S., Jeon, S. Y., Kim, K. O., et al. (2003). Consumers report preferences when they should not: A cross-cultural study. *Journal of Sensory Studies, 18*, 487–516.
Meilgaard, M., Civille, G. V., & Carr, B. T. (1991). *Sensory evaluation techniques* (2nd ed.). Boca Raton, FL: CRC Press.
Minoza-Gatchalian, M., & Divino-Brannan, G. (2009). *Sensory quality measurement. Statistical analysis of human responses*. Quezon City, Philippines: Quality Partners Company.
Moskowitz, H. R., Beckley, J. H., & Resurreccion, A. V. A. (2006). *Sensory and consumer research in food product design and development*. Oxford, UK: IFT Press, Blackwell Publishing.
Odesky, S. H. (1967). Handling the neutral vote in paired comparison product testing. *Journal of Marketing Research, 4*, 199–201.
O'Mahony, M., & Goldstein, L. (1987). Tasting successive salt and water stimuli: The roles of adaptation, variability in physical signal strength, learning, supra- and subadapting signal 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. A. P. D. E., & 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.
Prescott, J., Leslie, N., Kunst, M., & Kim, S. (2005). Estimating a "consumer rejection threshold" for Cork taint in white wine. *Food Quality and Preference, 16*, 345–349.
Resurreccion, A. V. A. (1998). *Consumer sensory testing for product development*. Gaithersburg, MD: Aspen Publications.
Stone, H., & Sidel, J. L. (2004). *Sensory evaluation practices* (3rd ed.). Oxford, UK: Elsevier Academic Press.
Sung, Y.-E., Lee, H.-S., O'Mahony, M., & Kim, K.-O. (2011). Paired preference tests: Use of placebo stimuli with liking and buying preferences. *Journal of Sensory Studies, 26*, 106–117.
Ullmann-Margalit, E., & Morganbesser, S. (1977). Picking and choosing. *Social Research, 44*, 757–785.
Villegas-Ruiz, X., Angulo, O., & O'Mahony, M. (2008). Paired preference 'placebo' tests with 'identical' stimuli: Does introducing graded preference responses affect the frequency of 'no preference' responses? *Journal of Sensory Studies, 23*, 439–449.
Weiss, B. H., O'Mahony, M., & Wichchukit, S. (2010). Various paired preference tests: Experimenter effect on "take home" choice. *Journal of Sensory Studies, 25*, 778–790.
Wichchukit, S., & O'Mahony, M. (2010a). A transfer of technology from engineering: Use of ROC curves from signal detection theory to investigate information processing in the brain during sensory difference testing. *Journal of Food Science, 75*, R183–R193.
Wichchukit, S., & O'Mahony, M. (2010b). Paired preference tests: 'Liking', 'buying' and 'take away' preferences. *Food Quality and Preference, 21*, 925–929.
Wichchukit, S., & O'Mahony, M. (2011). 'Liking', 'buying', 'choosing' and 'take away' preference tests for varying degrees of hedonic disparity. *Food Quality and Preference, 22*, 16–65.
Wichchukit, S., & O'Mahony, M. (2015). The 9-point hedonic scale and hedonic ranking in food science: Some reappraisals and alternatives. *Journal of the Science the Food and Agriculture, 95*, 2167–2178.
Wilke, K. D., Cochrane, C.-Y. C., & Chambers, E. (2006). Multiple preference tests can provide more information on consumer preferences. *Journal of Sensory Studies, 21*, 612–625.
Xia, Y., Zhong, F., & O'Mahony, M. (2015). Pairing detection of off-flavor in orange juice with preference tests. *Journal of Sensory Studies, 30*, 259–268.
Xia, Y., Zhong, F., & O'Mahony, M. (2016a). Paired preference testing: False preferences and disruptive protocols. *Food Science And Biotechnology, 25*, 1–10.
Xia, Y., Zhong, F., & O'Mahony, M. (2016b). Applying disruptive preference test protocols to increase the number of "no preference" responses in the placebo pair, using Chinese consumers. *Journal of Food Science, 81*, S2233–S2239.
Zhang, X., Halim, J., Wichchukit, S., O'Mahony, M., & Hautus, M. J. (2016). Paired preference tests: A signal detection based analysis with separate *d'* values for segmentation. *Journal of Sensory Studies, 31*, 481–491.