Received: 16 June 2015 | Revised: 22 August 2016 | Accepted: 4 October 2016

DOI 10.1111/joss.12238

ORIGINAL ARTICLE

WILEY Journal of Sensory Studies

Paired Preference Tests: A signal detection based analysis with separate d' values for segmentation

Xiaotian Zhang[1] | Jeremia Halim[1] | Sukanya Wichchukit[2,3] | Michael O'Mahony[1] | Michael J. Hautus[4]

[1]Department of Food Science and Technology, University of California, Davis, California

[2]Faculty of Engineering at Kamphaengsaen, Kasetsart University, Kamphaeng Saen Campus, Thailand

[3]Center of Excellence in Agricultural and Food Machinery, Kasetsart University, Thailand

[4]School of Psychology, The University of Auckland, Private Bag 92019, Auckland, New Zealand

Correspondence
M. J. Hautus, School of Psychology, The University of Auckland, Private Bag 92019, Auckland, New Zealand.
Email: m.hautus@auckland.ac.nz

Funding information
This research was partially supported by the Davis Sensory Science Foundation.

Abstract

Consumers gave graded preference responses to potato chips in a paired preference test. The graded responses were given to both the target pair under consideration and putatively identical "placebo" pairs of chips. From these data, a novel Signal Detection analysis was used. A model was developed giving a "magnitude of preference" distribution, for those consumers who preferred the first type of chip in the target pair and a second distribution for those who preferred the second type of chip. A second pair of distributions was generated for the two placebo pairs that had also been presented to the consumers. Using a signal detection paradigm, a value of d' was computed for each chip, representing the difference between the preference distribution for the target pair (Signal + Noise) and its corresponding placebo pair (Noise). The analysis has the advantage that these d' values are not distorted by the responses of those consumers who had reported preferences for the placebo pair. This advantage is not a feature of the regular computation of d' values based on the 2-AC test.

Practical applications

Paired preference tests are an important part of the measurement of consumer acceptance. Unfortunately, they are prone to response bias whereby consumers report preferences for putatively identical products; they are responding to the test conditions rather than the sensory properties of the products under assessment. Using identical products as a control, the effects of this on the results for the two different products to be assessed can be analyzed. There are various statistical approaches and models for solving this problem. This paper introduces an improved form of analysis based on signal detection/Thurstonian modeling. The method provides more meaningful information regarding segmentation.

1 | INTRODUCTION

As a casual test, the paired preference test (Lawless & Heymann, 2010; Resurreccion, 1998; Stone & Sidel, 2004) is a simple and convenient tool for preliminary guidance. It indicates the proportion of people tested, who prefer each product as well as the proportion who have no preference. Accompanying the test, questions regarding liking/disliking for each product should be included, to give further insight into the reasons for the measured preferences. For more formal testing, however, there are issues that need to be addressed. It is reasonable to assume that the goal of a formal preference test is to predict "real life" consumer behavior, whether it be choice behavior or buying behavior; it is not merely to predict behavior that is confined to the test situation. Wichchukit and O'Mahony (2011) taking a concept from Psychology, 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, an "operational preference." Thus, test preferences are intended to give insight into operational preferences.

One issue to be addressed is a tendency to report preferences when the stimuli are putatively identical. Ennis and Collins (1980) mailed two cigarettes (call them "A" and "B") to a large number of consumers' homes for comparison on a variety of attributes like: better flavor, easier draw, better aftertaste, slower burning, and so forth. Finally they were asked for their preferences and 40% reported preference for cigarette "A," 20% reported "No Preference" and 40% preference for "B." Yet, "A" had been taken from the initial part and "B" had been taken from the final part of the same production run. They were

essentially the same cigarette. Therefore, any preferences expressed for one or the other of the cigarettes would have been due to factors other than their sensory characteristics. This experiment was repeated with four different brands of cigarettes with consumer sample sizes ranging 412–488 (total 1,787). There was remarkable agreement between each brand.

Response to the putatively identical pair was hypothesized as a response to what were called "extraneous factors" in the testing situation. In other words, these preferences are not systematically related to the properties of the products in the test that are of relevance for an operational preference. This is because the sensory input elicited by the attributes of the putatively identical pair would be close enough to identical, to be deemed as being the same product (Calderon Rivera-Quintero, Xia, Angulo, & O'Mahony, , 2014; Sung, Lee, O'Mahony, & Kim, 2011). Therefore, it was hypothesized that because it is not logically possible to have an operational preference when two stimuli are identical, the response to such an identical pair could be taken as the response obtained when there is no preference. Accordingly, it was named the identicality norm (Ennis & Ennis, 2012a). The extraneous factors could be psychological in nature, like a preconception that the stimuli in a preference test must surely be different so that preference responses are expected and should be given. Various reasons for these preferences, as well as the relationship between "test preferences" "and operational preferences," have been discussed (Marchisano et al., 2003; Xia, Rivera-Quintero, Calderón, Zhong, & O'Mahony, 2014).

Ennis and Collins's (1980) 40-20-40 frequency distribution for putatively identical stimuli does not seem to be general. For reasons as yet unresolved, other authors (Alfaro-Rodriguez, Angulo, & O'Mahony, 2007, 2008; Alfaro-Rodríguez, O'Mahony, & Angulo, 2005; Alvarez-Coureaux, Aguilar, O'Mahony, & Angulo, 2010; Angulo, Okayama, Nakamura, Yuen, & O'Mahony, 2009; Calderón et al., 2015; Chapman, Grace-Martin, & Lawless, 2006; Chapman & Lawless, 2005; Sung et al., 2011; Kim, Lee, O'Mahony, & Kim, 2008; Marchisano et al., 2003; Xia et al., 2014) found different frequencies, the numbers varying with the products being tested, the experimental conditions, the types of consumers tested and the types and numbers of response options available in the test. The frequencies of reported "No Preferences" vary a great deal but most are in the range 20–35%. Yet, in nearly all cases, the majority of consumers indicate a preference rather than no preference.

The safest approach at the present time, would appear to be to require consumers to assess the target (different) pair and also the putatively identical (placebo) pair to give a measure of the effect of the extraneous factors. This was adopted and involved giving consumers two pairs to assess in a given test (Alfaro-Rodriguez et al., 2007, 2008; Alvarez-Coureaux et al., 2010; Kim et al., 2008; Sung et al., 2011). Using Chi-squared, the responses elicited by the target pair of (different) stimuli were not compared with the null hypothesis but were instead compared with the hypothesized identicality norm, derived from the putatively identical pair, to see whether they were significantly different. If they were, it could be concluded that the preference responses elicited by the sensory input from the attributes of the target pair were not completely a response to extraneous factors in the testing situation. At least some of the consumers were reacting to the sensory attributes of the stimuli; at least some of the consumers were indicating genuine preferences. Inspection of the distribution of responses could also determine whether the consumers were segmented into different groups. Because the putatively identical pair was being used in a way similar to a placebo in drug testing, it was called the "placebo" pair (Alfaro-Rodriguez et al., 2007).

For further sophistication, a Thurstonian/Signal Detection (Green & Swets, 1966; Lee and O'Mahony, 2004; O'Mahony and Rousseau, 2002) fundamental measure of perceived difference, d', was adopted. In this context, it was a single unitary measure representing an overall preference tendency of the consumers toward one item or the other (Angulo & O'Mahony, 2005). It can be conceptualized as giving a unitary numerical measure of how far the sample of consumers "leaned" toward one product or another. This is not always easy to discern from preference frequencies, if there is a substantial number of "no preference" choices. It also allows comparisons between studies using different experimental protocols, for example with and without the "no preference" option (Alfaro-Rodriguez et al., 2007; Angulo & O'Mahony, 2005). For d' values using tests without a "No Preference" response option, tables for the 2-AFC test (Ennis, 1993) can be used, while for tests with a "No Preference" response option, the computation for the 2-AC test is used (Braun, Rogeaux, Schneid, O'Mahony, & Rousseau, 2004). More importantly, for tests with a "No Preference" option, values of d' have been used in the comparisons between the target pair and the placebo pair, to determine whether they were significantly different (Alfaro-Rodriguez et al., 2007; Sung et al., 2011). Recently, there has been a renewed interest in the statistics associated with this computation (Christensen, Lee, & Brockhoff, 2012; Ennis & Ennis, 2012a,b; Christensen, Ennis, Ennis, & Brockhoff, 2014; Jesionka, Rousseau, & Ennis, 2014).

Despite the refinements for these computations, there is a fundamental problem. A significant difference only allows the conclusion that the preference responses elicited by the sensory input from the attributes of the target pair, were not completely a response to extraneous factors in the testing situation. It does not give information about the strength of the preference exhibited in the target pair. Using an analogy with Signal Detection Theory, the responses to the placebo pair can be considered as "Noise" in the system. Because those consumers who had reported preferences in the placebo pair were not eliminated from the experiment, the responses to the target pair represent "Signal + Noise." This is a standard situation for discrimination testing. Here, d' would represent the "preference signal strength," the strength of the preference, which would be obtained from the difference between the "Noise" and "Signal + Noise" distributions, as long as all the assumptions, like equal variance, and so forth, are fulfilled. Recently, Bi, Lee, and O'Mahony, (2015) published a model for this computation.

Nevertheless, this approach can be refined. Instead of using a single d' value to represent the signal strength of the preference in the target pair over the strength of the preference in the placebo pair, two d' values could be computed, to give more information regarding segmentation. It is possible to compute separate d' values for each product, each showing the difference between the product distribution in

TABLE 1 The response sheet with graded response options used for the simulation example and the response options for collection of preference data in the potato chip experiment

☐ I would only choose the one on the **left** and never choose the one on the **right**	☐ I would only choose the one on the **left** but I couldn't promise I would never choose the one on the **right**	☐ I would choose the one on the **left** much more than the one on the **right**	☐ I would tend to choose the one on the **left** a little more than the one on the **right**	☐ I would tend to choose each one roughly the same amount of time	☐ I would tend to choose the one on the **right** a little more than the one on the **left**	☐ I would choose the one on the **right** much more than the one on the **left**	☐ I would only choose the one on the **right** but I couldn't promise I would never choose the one on the **left**	☐ I would only choose the one on the **right** and never choose the one on the **left**
☐ I would not choose either of them								

the target pair and the product distribution in the placebo pair. In this way, the "preference signal strength" could be computed independently for each product.

The present paper outlines a method based on Signal Detection Theory that measures the signal strength of the preference for each product in a paired preference test. First, the computation will be illustrated with a simulation and second, experimental data will be analyzed.

2 | ILLUSTRATION OF THE METHOD AND ANALYSIS

For such a computation, consumers are required to give graded responses to represent their preferences or lack of preference. This approach has been used in previous preference testing protocols (Schwartz & Pratt 1956; Villegas-Ruiz, Angulo, & O'Mahony, 2008). A response sheet consistent with our example and employed in the experiment described in a later section is shown in Table 1. Note that, for simplicity, the verbal response options for preference are presented in terms of "likelihood to choose" so as to be more actionable. Strictly, there are several types of preference that can be measured. The common one is a "liking" preference, while other preferences like a "buying" preference, "choosing" preference and a "take away" preference have also been measured and compared (Sung et al., 2011; Weiss, O'Mahony, & Wichchukit, 2010; Wichchukit & O'Mahony, 2010). In this paper, we will refer to the choosing preferences measured here, simply as "preference." In Table 1, there are four levels of likelihood of choosing for each stimulus, with a neutral response option in the middle, giving a total of nine response options. This presents no surprises to the sensory community who are used to selecting verbal response options for the nine-point hedonic scale (Peryam & Giradot, 1952; Peryam & Pilgrim, 1957).

The responses to Table 1 are in terms of preferring a stimulus presented on the left or presented on the right. Let us assume that there are two products to be assessed for preference, in the test. We will call these P_1 and P_2. Because the position of these products will be counterbalanced when presented to the consumers, P_1 will sometimes be on the left and sometimes on the right. The same is true for P_2. Obviously, the data will need to be rearranged in terms of preferring P_1 or P_2 rather than the product on the left or the product on the right. Table 2 gives an example of such data for a simulation to illustrate the method.

In Table 2, the left hand column indicates which preference responses are presented in that row. For example, the row "Prefer P_1" indicates the preference responses for those who preferred P_1 (or had no preference) when it was compared with P_2. "Prefer P_2" gives the same data for those who preferred P_2 when compared with P_1. $<P_1$: $P_1>$ gives the preference responses for P_1 in the P_1 placebo pair, likewise for $<P_2$:$P_2>$.

The preference responses chosen for the simulation are given in terms of the number of consumers who responded to each of the five possible response options available for P_1 or P_2 that are derived from Table 1. For example, for "Prefer P_1," 11 consumers gave the maximum preference response: "I would only choose P_1 and never choose P_2,"

TABLE 2 Response frequencies for the simulation example based on the response options given in Table 1

	Choose roughly the same (No preference)	Choose a little more	Choose much more	Only choose this one, not promise never the other	Only choose this one, never the other	Total
Prefer P_1	12	16	23	6	11	68
Prefer P_2	12	11	15	17	9	64
$<P_1:P_1>$	23	23	8	5	1	60
$<P_2:P_2>$	17	21	6	9	7	60

while 16 consumers gave the minimum preference response: "I would tend to choose P_1 a little more than P_2." For "Prefer P_2," the corresponding numbers of consumers are 9 and 11. For the placebo pair, $<P_1:P_1>$, as might be expected in reality, only 1 consumer chose the maximum preference option of "only choosing one product" while 23 chose the minimum preference option of "choosing one product a little more." For $<P_2:P_2>$, the corresponding numbers of consumers are 7 and 21.

For the signal detection analysis, the preference measure is represented in Table 2 by the distribution of the number of consumers in a sample, choosing the various response options: maximum preference option for cells on the right and minimum preference option for cells on the left. The simulation considered a sample of 120 consumers. Therefore, with counterbalancing, 60 consumers would be presented with the $<P_1:P_1>$ placebo pair and 60 with $<P_2:P_2>$. These totals are given in the right hand column of Table 2. For the target pairs (P_1 vs. P_2), there were also 120 presentations, given in counterbalanced order (60 of each order). It might be expected that as with the placebo pairs, the total for "Prefer P_1" and "Prefer P_2" should also be 60. Yet, this is not so; they add up to 64 and 68, respectively. One reason is because there is "doubling" in the "No Preference" column. In this simulation, when P_1 was presented on the left and P_2 on the right, 6 consumers chose the "No Preference" option. When P_2 was presented on the left and P_1 on the right, 6 consumers also chose the "No Preference" option, giving a total of 12 consumers choosing "No Preference." This number is indicated in the "Prefer P_1" row. However, the same 12 consumers are also represented in the "Prefer P_2" row. Therefore, they have been represented twice. One way of dealing with this is to put half in the "Prefer P_1" row and half in the "Prefer P_2" row, giving totals of 66. These would be the totals if all the other preference responses had been divided equally between "Prefer P_1" and "Prefer P_2" However, they were not. "Prefer P_1" had 56 responses and "Prefer P_2" had 52, giving totals of 68 and 64, respectively.

The model for this is shown in Figure 1, which represents the population from which the consumer sample was drawn. Spread along the horizontal axis of this figure, there are four preference distributions. One is for those who preferred P_1 in the target pair and one for those who preferred P_2 as well as one for each of the placebo pairs, representing preferences expressed as a response to extraneous factors. The model requires a reference (zero) point on the axis. For this, the mean of the left hand $<P_1:P_1>$ placebo distribution is arbitrarily set as zero.

The four solid vertical lines are preference boundaries, which represent the divisions between the verbal response categories in Table 1. The spacing between these boundaries is governed by the number of consumers who chose each category.

This model can then be fitted to the data in Table 2 with the use of standard techniques. Here maximum likelihood estimation was employed. The model is similar to that described by Dorfman and Alf (1969) and their Equation 3 was used, extended to four distributions, to define the log likelihood, which was then maximized by iteratively adjusting the available parameters. This can be readily accomplished in Excel using the inbuilt Solver tool. To estimate standard errors for each fitted parameter, Mathematica™ (Wolfram Research Inc., IL) can be employed.

Table 3 shows the best-fitting parameter estimates that correspond to the model illustrated in Figure 1. Each fitted parameter estimate has an associated standard error (*SE*) useful for undertaking statistical comparisons. Note that the mean of the $<P_1:P_1>$ distribution was fixed at zero, and hence there is no standard error associated with this parameter.

The goodness-of-fit of the model to these simulated data can be assessed using the *G*-statistic. In this case $G^2 = 19.63$. Larger numbers indicate poorer fits. An estimate of the probability that the observed data could have arisen from the model, given that the model is correct,

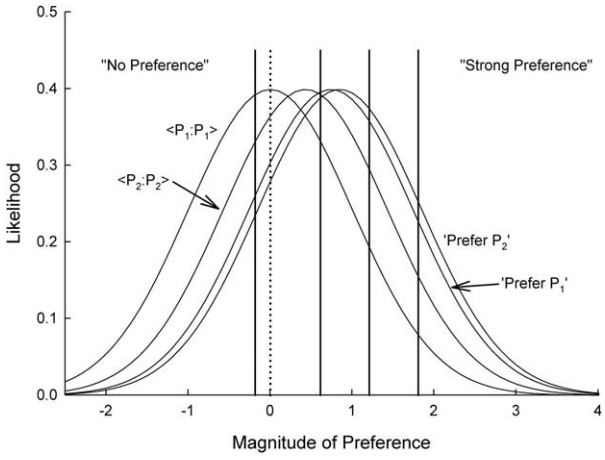

FIGURE 1 Preference distributions fitted to the simulated data in Table 2. The four distributions are labeled. The four solid vertical lines are the preference boundaries (z_1. z_4). All parameter values are provided in Table 3

TABLE 3 Best-fitting parameter estimates that correspond to the model illustrated in Figure 1. Each fitted parameter estimate has an associated standard error (*SE*) useful for undertaking statistical comparisons. Note that the mean of the $<P_1:P_1>$ distribution is fixed at zero, and hence there is no *SE* associated with this parameter

Distribution	Mean	*SE*	Preference Boundaries (z_1.z_4)	*SE*
$<P_1:P_1>$	0	–	−0.186	0.145
$<P_2:P_2>$	0.426	0.196	0.610	0.147
Prefer P_1	0.738	0.190	1.206	0.156
Prefer P_2	0.845	0.194	1.802	0.172

can be obtained from the *G*-statistic. Asymptotically, the *G*-statistic is distributed as chi-square. The number of degrees of freedom of the chi-square distribution is equal to the number of degrees of freedom from the data less the number of estimated parameters. Each row in Table 2 contains five frequencies and their total, which is fixed. Four of the frequencies can vary, and the fifth is that frequency needed to sum all five to the fixed total. Hence there are four degrees of freedom associated with each row, and four rows give a total of 16 degrees of freedom. The number of fitted parameters is 7 (3 Means and 4 preference boundaries, see Table 3). Hence the *G*-Statistic is distributed as chi-square with $16 - 7 = 9$ degrees of freedom and $P(\chi^2(9) > 19.63) = 0.020$. For these simulated data, if the best-fitting model is correct, there is about a 2% chance of obtaining the data we observed, or poorer-fitting data. As far as models go, that is not bad. A probability less than .001 is often suggested as appropriate for rejecting a model (Bentler & Bonett, 1980; Press, Flannery, Teukolsky, & Vetterling, 1992).

Assessing differences between the distributions is simplified by them all having the same variance. This equal variance also indicates that distances along the *x*-axis can be interpreted as values of d'. Differences between the distributions can be assessed by comparing their means (see Table 3). The standard approach used is to undertake *z*-tests using Equation 1a to test two fitted means and to use Equation 1b to test the difference between a fitted mean and zero, which in this model is equivalent to testing for a difference from the $<P_1:P_1>$ placebo distribution.

$$(a)\ z = \frac{\mu_2 - \mu_1}{\sqrt{SE_1^2 + SE_2^2}} \quad \text{or} \quad (b)\ z = \frac{\mu}{SE} \tag{1}$$

The resulting value of *z* can then be looked up in normal distribution tables to obtain the associated level of significance. If $\mu_2 > \mu_1$, a value of $z > 1.96$ corresponds to a significant result at the alpha = 0.05 level.

First, we shall test whether the distribution means are significantly different from zero, the mean of the $<P_1:P_1>$ distribution. Consider the *z*-values based on Equation 1b associated with the three fitted means and their standard errors (*SE*). For $<P_2:P_2>$, $z = 0.426/0.196 = 2.17$; for "Prefer P_1," $z = 0.738/0.190 = 3.88$; and for "Prefer P_2," $z = 0.845/0.194 = 4.36$. Note that all *z*-values are above 1.96 and so all means are significantly different from zero. Of particular note here is that the means of the two placebo distributions are significantly different.

A significant difference between a preference distribution and its associated placebo distribution, equivalent to the identicality norm, suggests a true consumer preference for the respective product. For example, if the mean of the "Prefer P_1" distribution is significantly greater than that for the $<P_1:P_1>$ placebo distribution ($\mu = 0$) then the group of consumers who said they preferred product P_1 have a significantly larger magnitude of preference than that exhibited by consumers for the $<P_1:P_1>$ placebo pair. This particular comparison was made above (0.738/0.190: $z = 3.88$) indicating that those who preferred P_1 over P_2 did so with a magnitude of preference that exceeded that for the placebo distribution. Equation 1a is used to test the "Prefer P_2" distribution against the P_2 placebo distribution, $<P_2:P_2>$; $z = (0.845 - 0.426)/\sqrt{(0.194^2 + 0.196^2)} = 1.519$. This result is not significant ($z < 1.96$). Thus, there is evidence for a significant strength of preference for those who preferred P_1, but not for those who preferred P_2. It is possible for the group with the highest mean preference magnitude to not show a significant preference, whilst a group with a lower mean preference magnitude does show a significant preference. The two preference groups can be compared directly: $z = (0.845 - 0.738)/\sqrt{(0.194^2 + 0.190^2)} = 0.394$. The result is not significant ($z < 1.96$). Thus, these two groups exhibit the same level of preference magnitude.

In simple language, because the means of the "Prefer P_1" and "Prefer P_2" distributions are not significantly different from each other, it would be concluded that if only the target pair was assessed, preference for P_1 and P_2 would not be significantly different. It should be remembered, however, that the target pair includes consumers who are sufficiently biased to report preferences when assessing the placebo pair. In signal detection terms "Prefer P_1" and "Prefer P_2" represent "Signal + Noise" values for the preference. The "Noise" levels are represented by the placebo pairs, of which the $<P_2:P_2>$ distribution has a significantly higher mean than $<P_1:P_1>$. Just as the true signal strength is given by the difference between the means of the "Signal + Noise" and "Noise" distributions, the true strength of preference for the two products is given by the difference between the means of the "Prefer P_1" distribution and the $<P_1:P_1>$ distribution and between the distributions: "Prefer P_2" and $<P_2:P_2>$. The difference between the "Prefer P_1" distribution and the $<P_1:P_1>$ distribution is significant, indicating a real preference for P_1 when compared with P_2. Yet, the difference between "Prefer P_2" and $<P_2:P_2>$ is not significant, indicating that any preference for P_2 when compared with P_1 is no greater than the "Noise" level; there is no significant preference for P_2. This is the strength of the method. Should preferences be deduced from the target pair alone, the conclusion would be that there is no overall preference for P_1 or P_2. Yet, if the true preference strengths are measured in relation to their placebos, there is a significant preference for P_1 and only a non-significant preference for P_2.

In Figure 1, the four preference boundaries are located in a manner that best fits the data, in terms of proportion of consumers surrounding each boundary. Hence they can be used to make statements about the population of consumers from which participants were representatively

sampled. For example, the area under each of the distributions to the left of the leftmost boundary, can be taken as the model's prediction of the proportion of consumers in the population who would respond with "No Preference."

Free software has been developed (Hautus, 2012) for these calculations. For this, the measured response frequencies can be entered into a version of Table 1, which is adjusted to name the actual products used in the test, rather than "left" and "right." For this table, let the products be named "A" and "B." With a click, data will appear from which d′ values can be computed. There will be 4 titles labeled "Boundary," below which will be values z_1, z_2, z_3, and z_4, below which will be numbers of the points on the axis representing the positions of the four boundaries. To the right of these, will be values with the label "Mean." These represent the means for the distributions. One of the placebo means (AA or BB) will have been chosen as the zero point on the axis. The second "Mean" (AA or BB) will have its position on the axis written below. The two following means will have the positions for "Prefer A" and "Prefer B." It is then a simple matter to obtain the appropriate d' values by subtraction.

This method of data analysis was applied to a two-part preference test involving potato chips. The preferences were expressed in terms of choice when pairs of chips were presented to the consumers. This is more actionable than simply asking for preference.

3 | MATERIALS AND METHODS

3.1 | Consumers

For Part A, a total of 299 consumers of potato chips (127M, 172F, age range 7–61 year, mean: 22.6 year), students, staff, and friends from the University of California, Davis, were intercepted in a campus dining room. They were tested at a table set up for the experiment, facing the experimenter. For Part B, a total of 247 consumers of potato chips (103M, 144F, age range 7–57 year, mean: 22.5 year) were sampled from consumers who were tested in Part A. Part B was performed after Part A in a single experimental session. Due to time constraints, not all consumers remained for Part B.

3.2 | Stimuli

For Part A, the stimuli comprised of two types of potato chip: "Honey Barbecue" (H) and "Cheddar and Sour Cream" (C). For Part B, they were comprised of "Barbecue" (B) and "Sour Cream and Onion" (S) (Frito-Lay, Inc., Plano, TX). The chips were presented in 2 oz (59 ml) Solo plastic cups (Dart Container Corporation, Highland Park, IL) five to six pieces a cup. The chips had easily distinguishable flavors and could be recognized easily by their different appearance. Consumers were allowed to taste more if required.

3.3 | Procedure

Consumers were tested individually. After establishing rapport and collecting demographic details, the experimenter instructed the consumers in the experimental procedure and only began the experiment when the consumers had thoroughly understood their task. The experimenters were careful to make no implications about how consumers were meant to respond. They also made sure that the consumers knew that they were not being tested for their tasting skill in any way; they were merely being asked for their opinions.

For Part A, the consumers were presented with two sets of chips. The first set, the target pair, consisted of a cup of "Honey Barbecue" (H) and a cup of "Cheddar & Sour cream" (C). They were told to indicate which chips they would choose if presented with both on a tray or whether they liked them just about the same and would choose either or both. Consumers were allowed to sample as many chips as desired from each cup. They gave their preference responses, by choosing (pointing and/or speaking) one of the response options on a response sheet (see Table 1). The response sheet was presented on a cardboard strip measuring 48 cm × 10.5 cm (18.9 in × 4.1 in). All writing was in black (font type: "Calibri," font size: 18) on a white background. To preserve reality, water rinses were not presented.

Immediately after this, the consumers were presented with the placebo pair. This consisted of two cups of putatively identical chips (either "Honey Barbecue" <H:H> or "Cheddar Cream & Onion" <S:S>). The placebo pair presented to consumers depended on which chips were preferred from the target pair. For consumers who chose "Honey Barbecue," the appropriate placebo pair consisted of two cups of "Honey Barbecue." For those who chose "Cheddar & Sour Cream," the appropriate placebo pair consisted of two cups of "Cheddar & Sour Cream." For the few (<10%) consumers who had no preference with the target pair, the first tasted chip determined the placebo pair. For Part B, the same procedure was used for the "Barbecue" and the "Sour Cream and Onion" chips.

It should be noted that the order of tasting for the target and placebo pairs was not counterbalanced. This was so that the appropriate placebo pair could be chosen, to correspond with the choice made in the target pair. Also, it meant that both placebo pairs were tested in the same context, namely after a target pair. This would not have happened with counterbalancing. Finally, the placebo pair ("Noise") was tasted immediately after the target pair ("Signal + Noise"). The closeness in time was to give a better chance of the noise levels in the placebo pair and in the target pair being close, so as to give the best estimate of the signal strength of the preference signal per se.

The order of presentation of the chips within the target pair was counterbalanced. For part A, half the consumers tasted "Honey Barbecue" first while the other half tasted "Cheddar & Sour Cream" first. Likewise for "Barbecue" and "Sour Cream and Onion" flavors in part B. The consumers always sampled the chips in the left hand cup first. The total experimental session time ranged 5–20 min.

4 | RESULTS AND DISCUSSION

For Part A, the responses of the 299 consumers to tasting "Honey Barbecue" (H) and "Cheddar and Sour Cream" chips (C) were arranged according to the various response options displayed on the response sheet (Table 1). These are given in the top part of Table 4 which follows the format

TABLE 4 Response frequencies based on the response options in Table 1, for target pairs and placebo pairs for "Honey Barbecue" and "Cheddar and Sour Cream" chips presented in Part A and sour "Cream and Onion" and "Barbecue" chips presented in Part B

	Choose roughly the same (No preference)	Choose a little more	Choose much more	Only choose this one, not promise never the other	Only choose this one, never the other	Total
Part A, a total of 299 consumers. Chips: Honey Barbecue (H), Cheddar & Sour Cream (C)						
Prefer H	23	59	57	32	21	192
Prefer C	23	41	46	12	8	130
<H:H>	90	54	24	7	2	177
<C:C>	72	26	15	8	1	122
Part B, a total of 247 consumers. Chips: Sour Cream & Onion (S), Barbecue (B)						
Prefer S	26	43	45	18	19	151
Prefer B	26	25	48	16	7	122
<S:S>	79	34	20	3	3	139
<B:B>	53	24	19	7	5	108

used for Table 2. The table consists of four rows. The top two rows (Prefer H" and "Prefer C") give the responses to the target pairs. Here, because the 23 "No Preference" responses were common to both sets of data, they are necessarily included twice. The bottom two rows (<H:H> and <C:C>) give data for the placebo pairs. In the simulation, half the consumers were given one placebo and half the other. Accordingly, the totals for each placebo pair were the same (60). However in this experiment, the numbers of consumers performing each placebo were not the same, they depended on which product was preferred in the target pair. Accordingly, the totals were not the same. In the lower half of the table, the same data are displayed for the 254 consumers who tasted "Barbecue" (B) and "Sour Cream and Onion" (S) chips, tested in Part B.

4.1 | Analysis for part A

As in the simulation, the model was fitted to the data in Table 4 and the parameters of the resulting model are displayed in Table 5. For Part A, the mean for <C:C> was set to zero.[1] The other means were: <H:H> = 0.101, "Prefer H" = 1.208, and "Prefer C" = 0.953. The locations of the preference boundaries (the vertical lines) and the standard errors of all parameters are also provided.

The *G*-statistic, which indicates the goodness-of-fit of the model to these preference data, is 12.92. These data have the same basic structure as those described in the simulation example, and hence they also have 9 degrees of freedom. $P(\chi^2(9) > 12.92) = 0.116$, which is insufficient to reject the model, even at the conventional 0.05 level of significance. Even if two more degrees of freedom were dropped, taking into account the lack of independence between the two "no preference" categories, with a frequency of 23 in the top-left of Table 4, the model still provides an acceptable fit; $P(\chi^2(7) > 12.92) = 0.074$. All-in-all, this is a model that accounts well for these data. The model is illustrated in Figure 2.

Having obtained the parameters for the model and an illustration, the next task was to determine whether there were significant differences in preference for "Honey Barbecue" (H) and "Cheddar and Sour Cream" (C) chips. The first step was to determine whether the three distributions were significantly different from zero (the mean for <C:C>). The second step was to inspect the two preference distributions ("Prefer H," "Prefer C"). This was to determine whether they would have been taken as indicating a preference should placebo distributions be ignored. The third step was to consider the placebo distributions. The true strength of the preference for each product was then determined

TABLE 5 Best fitting parameter estimates that correspond to the models illustrated in Figures 2 and 3. Each fitted parameter estimate has an associated standard error (*SE*) useful for undertaking statistical comparisons. Note that the means of the <C:C> and <S:S> distributions are fixed at zero for Parts A and B, and hence have no associated *SE* values

Distribution	Mean	*SE*	Preference Boundaries (z_1.z_4)	*SE*
Part A, $G^2 = 12.92$, $p = .116$ Chips: Honey Barbecue (H), Cheddar & Sour Cream (C)				
<C:C>	0	–	0.117	0.106
<H:H>	0.101	0.134	0.978	0.110
Prefer H	1.208	0.131	1.795	0.119
Prefer C	0.953	0.140	2.431	0.136
Part B, $G^2 = 12.67$, $p = .178$ Chips: Sour Cream & Onion (S), Barbecue (B)				
<S:S>	0	–	0.172	0.099
<B:B>	0.288	0.145	0.860	0.104
Prefer S	1.042	0.132	1.722	0.116
Prefer B	0.933	0.138	2.224	0.130

[1]The software can be set to fix the mean of either of the placebo distributions to zero. Choosing the distribution farthest to the left in the fitted model may lead to easier interpretation of the results.

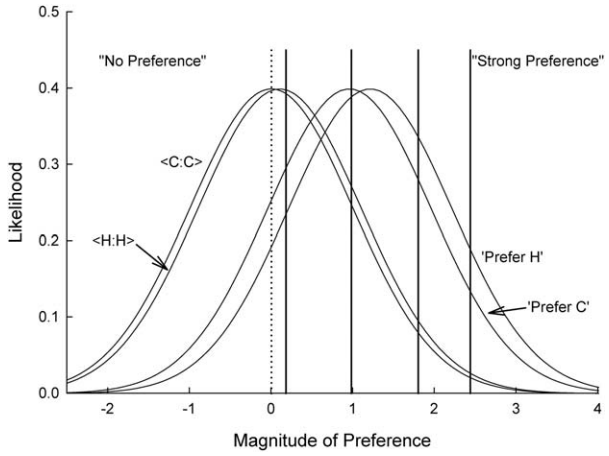

FIGURE 2 Preference distributions fitted to responses to chips, representing preference for "Honey Barbecue" chips (H) or "Cheddar and Sour Cream" chips (C) and responses to the <H:H> and <C:C> placebo pairs in Part A

by examining the difference between each of the "Prefer" distributions and their appropriate placebos.

Before considering the first step, it should be noted that a more general model was also fitted to the data. This model allowed the four distributions to have unequal variances, rather than the equal variances illustrated in Figure 2. The model with equal variances can be considered a special case of the model with unequal variances. Tests of the significance of differences between models related in this manner (nested models) to account for the data, are available (e.g., Kendall & Stewart, 1979). In this context, minus twice the difference between the Log Likelihood (LL) of the data for each model ($-2\ \Delta$LL) is the recommended test statistic, and it is distributed as chi-square with three degrees of freedom, because the unequal-variance model has three more parameters (the three variances) than the equal-variance model.[2] The probability associated with $-2\ \Delta$LL can be used to assess whether the additional parameters provide a significantly better fit to the data. For these data, $-2\ \Delta\text{LL} = 6.59$ and $P(\chi^2(3) > 6.59) = 0.086$. Thus, the unequal-variance model does not provide a statistically significant improvement in fit to these data.

For the first step, the four distributions in Figure 2 were examined to determine whether the mean of each preference group was significantly different from zero (Equation 1b). For the "Prefer H" group $z = 1.208/0.131 = 9.22$ ($p < .001$). For the "Prefer C" group $z = 0.953/0.140 = 6.81$ ($p < .001$). Thus, it can be concluded the means of both groups are significantly different from zero, indicating a statistically significant magnitude of preference (preference signal strength) in both cases.

The two placebo distributions in Figure 2 are close to each other, with a difference in means of 0.101. As noted earlier, the difference between the means can be interpreted as a value of d'. The significance of the difference, which is equivalent to a test of whether d' is greater than zero, can be assessed using Equation 1b. In this case, $z = 0.101/0.134 = 0.75$. A two-tailed test based on this z-value yields a probability of $p = .453$, indicating that the two placebo means are not significantly different. This suggests that the responses to extraneous factors in the testing situation are, in this case, the same for both placebo pairs. This might not necessarily always be the case.

The second step was to examine the two distributions for those who preferred one type of chip over the other. The distributions have a difference in means of 0.255. The null probability, associated with a z-test using Equation 1a for this difference is $p = 0.184$. This lack of a significant difference does not mean that the consumers did not have preferences. What it does mean is that the pattern of preference responses across the response sheet was almost the same for those who preferred chip H and those who preferred chip C. This is not always the case.

For the third step, the differences between each preference group and its respective placebo distribution are considered. The difference between the "Prefer C" distribution and <C:C> (zero) was calculated in step one, indicating a significant difference ($p < .001$). Equation 1a is used to test the "Prefer H" distribution against the <H:H> placebo distribution: $z = (1.208–0.101)/\sqrt{(0.131^2 + 0.134^2)} = 5.907$. This result is significant ($p < .001$). Thus there is evidence for a significant strength in preference for those who preferred chip C ($d' = 0.953$) and those who preferred chip H ($d' = 1.208 - 0.101 = 1.107$). However, the difference between these two d' values is not significant ($z = 1.107 - 0.953)/\sqrt{((0.131^2 + 0.134^2) + 0.140^2)} = 0.658$) at $p = .510$.

It can be concluded, that because the "Prefer H" and "Prefer C" distributions were significantly different from their respective placebo distributions, the sample of consumers was segmented into two groups, each with a preference that was greater than chance. When the noise level was not taken into account, the difference between the strengths of preference for the "Honey Barbecue" and the "Cheddar and Sour Cream" chips was not significant. When the noise level was taken into account the result was the same. However, the significance level changed (from 0.184 to 0.510) indicating a greater confidence in not rejecting the null hypothesis.

As an alternative analysis, if the difference between the means for the placebo distributions is not significant as in this case ($z = 0.101/0.134 = 0.75$), it would seem reasonable to use the midpoint between the two distributions as a reference point with which to determine degree of preference. One reason for this approach is that if the two placebo distributions do not differ significantly from each other, they can be considered as a single placebo distribution, and a good estimate of the mean of the combined distribution will be the average of the means of each distribution. That reference point, in this case, is $(0 + 0.101)/2 = 0.050$ and the associated SE is $\sqrt{(0.5^2 \times 0.134^2)} = 0.067$, which is half that for the <H:H> distribution alone. The distributions for the preference groups are closer to this reference point than from the original reference value of zero. For the "Prefer H" and "Prefer C" demographic groups, the re-referenced d' values are $1.208 - 0.050 = 1.158$ ($SE = \sqrt{(0.067^2 + 0.131^2)} = 0.147$) and $0.953 - 0.050 = 0.903$ ($SE = \sqrt{(0.067^2 + 0.140^2)} = 0.155$), respectively. These values are still significantly different from zero ($z = 1.158/0.147 = 7.88$ and $z = 0.903/0.155 = 5.83$, respectively). Note that the

[2]Note that one of the four variances is fixed at 1, in the same way that one of the four means is fixed at 0.

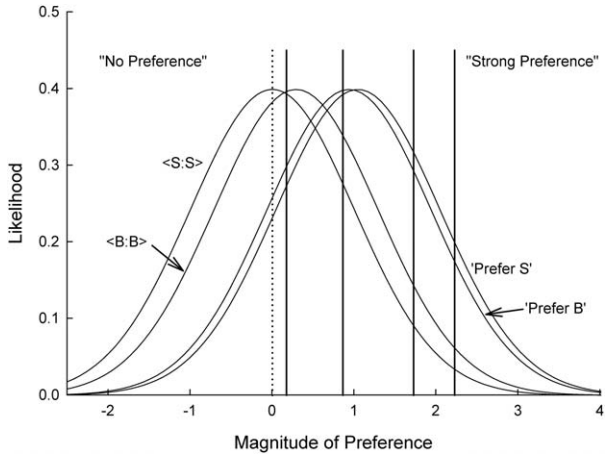

FIGURE 3 Preference distributions fitted to responses to chips, representing preference for "Sour Cream and Onion" chips (S) or "Barbecue" chips (B) and responses to the <S:S> and <B:B> placebo pairs in Part B

<C:C> distribution was to the left of the <H:H> distribution in this case, thus leading to a decrease in d', this will not always be the case. This difference is not significant with $z = (1.158 - 0.903)/\sqrt{(0.147^2 + 0.155^2)} = 1.194$. With this analysis, the same conclusions may be drawn.

4.2 | Analysis for part B

Frequencies of each response for the 247 consumers when tasting "Barbecue" (B) and "Sour Cream and Onion" (S) chips are given in the bottom part of Table 4. As in Part A, the model was fitted to these data and the best-fitting parameters are displayed in Table 5. The resulting model is illustrated in Figure 3. For these data, the mean for <S:S> was set to zero. The other means are: <B:B> = 0.288, "Prefer S" = 1.042 and "Prefer B" = 0.933. All parameters and their standard errors are provided in Table 5.

The goodness-of-fit of the model, is given by $G^2 = 12.67$, again with 9 degrees of freedom, yielding $P(\chi^2(9) > 12.67) = .178$. Even if two more degrees of freedom were dropped, for the lack of independence between the two "no preference" categories, the model still provides an acceptable fit; $P(\chi^2(7) > 12.67) = .081$.

As with Part A, an unequal-variance version of the model was also fitted to the data. For these data, $-2\,\Delta LL = 2.71$ and $P(\chi^2(3) > 2.71) = 0.439$. Thus, the unequal-variance model did not provide a statistically significant improvement in fit to these data.

The data were analyzed using the same steps as in Part A. For determining differences from zero, the "Prefer B" group has $z = 0.933/0.138 = 6.76$ ($p < .001$). For the "Prefer S" group $z = 1.042/0.132 = 7.89$ ($p < .001$). Thus, the means of both groups are significantly different from zero. The two placebo distributions have a difference in means of 0.288. In this case, $z = 0.288/0.145 = 1.98$, indicating that the two placebo means were significantly different ($p = .048$). Unlike for Part A, this suggests that the responses to extraneous factors in the testing situation were slightly different for both placebo pairs. Given that the difference between the placebo distributions was significant, we do not attempt the alternative analysis, using as reference the midpoint between the two distributions.

The question becomes one of why there were more preferences given for the <B:B> placebo than for the <S:S> placebo. In other words, why there were more preference responses when tasting a "Barbecue" placebo pair than when tasting a "Sour Cream and Onion" placebo. It could be hypothesized that the "Barbecue" chips elicited a sensory input that was more prone to sequence effects than the "Sour Cream and Onion." Trigeminal input from the barbecue flavor could have caused differences in perception when one chip was tasted after the other. Adaptation effects for a suitably strong taste could also have caused perceptual differences due to interactions in the oral cavity. This explanation would come under the heading of the "Input Hypothesis" discussed in Xia, Zhong, and O'Mahony, (2016).

Considering just the distribution for "Prefer S" and "Prefer B," the former was very slightly stronger than the latter. The two distributions have a non-significant difference in means of 0.109 resulting in a z-value of 0.57 ($p = .568$), as is to be expected from Figure 3.

The next step was to compare the preference distributions with their appropriate "Noise" levels. The comparison between those who preferred chip S and the <S:S> placebo, indicated that those who preferred chip S did so with a magnitude of preference that significantly exceeded that for its placebo distribution ($p < .001$). The corresponding test for the "Prefer B" distribution against the <B:B> placebo distribution yielded: $z = (0.933 - 0.288)/\sqrt{(0.138^2 + 0.145^2)} = 3.22$. This result was significant ($p = .0012$). Thus there was evidence for a significant strength in preference for those who preferred chip S ($d' = 1.042$) and for those who preferred chip B ($d' = 0.933 - 0.288 = 0.645$). For this analysis, d' for chip S was greater than that for chip B. Traditionally, the difference would not be considered significant ($z = (1.042 - 0.645)/\sqrt{(0.132^2 + (0.138^2 + 0.145^2))} = 1.656$) with $p = .098$. However, working at $\alpha = 0.10$, it would be regarded as significant.

In summary, for both Part A and B, these results indicate that in both cases there were two segmented groups of consumers of roughly the same size. These groups had opposite preferences, which were not significantly different in their strength. This was apparent from inspecting just their preference distributions. Yet, when the noise levels represented by the placebo pairs were taken into account, the sureness regarding the lack of difference in preference strength was changed. For Part A, the sureness was increased. For Part B, the sureness was decreased; in fact, the preference for "Sour Cream and Onion" over the "Barbecue" chips would be regarded as significant at $p < .10$.

5 | CONCLUDING REMARKS

As described in the Introduction, when d' values have been used as a measure of strength of preference, the approach has been different. A d' is calculated giving the overall preference or lack of preference for one of the products in the target pair, based on the 2-AC methodology. The same is done for a single overall placebo pair (both possible pairs combined). The latter is regarded as the "No Preference" or "Noise" level in the experiment. For an estimated infinite number of placebo pairs, it is

called the "Identicality Norm." This is compared with the d' value for the target pair of products, and should these two d' values be significantly different, it is concluded that there is a significant preference present in the target pair (Alfaro-Rodriguez et al., 2007; Ennis & Ennis 2012a; Sung et al., 2011). The present paper refines this Signal Detection approach by requiring consumers to use separate placebo pairs, depending on their product preference with the target pair. This provides a separate d', representing the strength of preference for each product, making any differences due to segmentation more easily assessed.

Besides the present equal variance model, it was important to compare whether an unequal variance model gave a better fit. For difference tests, where the stimuli are confusable, an equal variance assumption is reasonable. Yet, for preference judgements, such an assumption needs to be checked. Should the two products be at separate ends of a positive/negative preference dimension, it could be argued that they might have different ranges of acceptance and thus different variances. Yet, should they be close on this dimension, it could be argued that the variance, whether large or small could be similar. In the present experiment their variances were similar

Interestingly, from Table 4, it can be seen that with graded responses the "No Preference" option is most commonly chosen for placebo pairs. The frequency ranges 49–69%. Yet, as mentioned in the "Introduction," "No Preference" responses from previous research usually range 20–35%. In this study, the target pair was always presented before the placebo, which is not true for the earlier studies. This could have given consumers an idea of the size of differences between the two target stimuli so that in contrast, the placebo pair would be more likely to be judged as the same. There is always the logical possibility that the unusual graded response protocol of the present study might have had an effect, although the reason why is not obvious.

The second most commonly chosen option for placebo pairs is "Choosing one product a little more". Although this is not surprising, it is worth noting because this may act as a clue when considering differences between "real" and "false" preferences, when only target pairs are presented to consumers.

It can be argued that if those consumers who indicate their tendency to respond to extraneous factors, by responding with preferences to the stimuli in the placebo pair, are included in the sample of consumers assessing the target pair, it is not a matter for concern. It is argued that it will still be possible to see which product is preferred. Conversely, the extent of preference will not be clear. Consider a sample of consumers with a 70:30 preference ratio for a given product in the target pair. Then consider adding to the sample, the same number of consumers, who reported random, hence equal preferences for the stimuli in the placebo pair. If they continued to respond randomly, the preference ratio would change to 60:40. Such a change could elicit different marketing decisions. However, it could also be argued that such consumers, when confronted with two obviously different products in the target pair, would no longer have a tendency to respond to extraneous factors. This may be true, but the proportion of consumers who did this would be unknown and the extent of distortion in the results could not be ascertained.

The present study is but one attempt to account for the effects of "noise" in the system. Other approaches have been used. One such approach is to use so-called "disruptive protocols," by altering the experimental conditions so that the number of consumers who respond to extraneous factors and consequently report preferences for the placebo pair is drastically reduced (Calderón et al., 2015; Xia et al., 2014).

ACKNOWLEDGMENT

This research was partially supported by the Davis Sensory Science Foundation.

REFERENCES

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

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, M., & O'Mahony, M. (2008). Paired preference tests: '50:50' and 'Alternating' no preferences. *Journal of Sensory Studies*, *23*, 765–779.

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 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.

Angulo, O., Okayama, T., Nakamura, T., Yuen, R., & O'Mahony, M. (2009). Use of purchase preference options to increase 'no preference' frequencies in placebo preference tests. *Journal of Sensory Studies*, *24*, 258–268.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606.

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

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., Rivera-Quintero, A., Xia, Y., Angulo, O., & O'Mahony, M. (2015). The triadic preference test. *Food Quality and Preference*, *39*, 8–15.

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 no-preference option in dairy product testing. *Journal of Sensory Studies*, *20*, 454–468.

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

Christensen, R. H. B., Ennis, J. M., Ennis, D. M., & Brockhoff, P. B. (2014). Paired preference data with a no-preference option––

Statistical tests for comparison with placebo data. *Food Quality and Preference*, *32*, 48–55.

Dorfman, D. D., & Alf, E. (1969). Maximum-likelihood estimation of parameters in signal-detection theory and determination of confidence intervals––Rating-method data. *Journal of Mathematical Psychology*, *6*, 487–496.

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

Ennis, D. M., & Collins, J. (1980). *The distinction between discrimination and splitting in paired testing* (Report # 80-233). Richmond, Virginia: Philip Morris Research Center, pp. 50.

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.

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

Hautus, M. J. (2012). SDT Assistant. (Version 1.0) [Software] Retrieved from http://hautus.org

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.

Kendall, M. G., & Stewart, A. (1979). *The advanced theory of statistics*, (4th ed.). New York: Macmillan Publishing.

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.

Lawless, H. T., & Heymann, H. (2010). *Sensory evaluation of food principles and practices*, (2nd ed.). New York: Springer.

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

Marchisano, C., Lim, J., Cho, H.-S., Suh, D.-S., Jeon, S.-Y., Kim, K. O., & O'Mahony, M. (2003). Consumers report preference when they should not: A cross-cultural study. *Journal of Sensory Studies*, *18*, 847–516.

Peryam, D. R., & Girardot, N. F. (1952). Advanced taste-test method. *Food Engineering*, *24*, 58–61.

Peryam, D. R., & Pilgrim, F. J. (1957). Hedonic scale method for measuring food preferences. *Food Technology*, *11*, 9–14.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical recipes in C: The art of scientific computing*, (2nd ed.). New York: Cambridge University Press.

Resurreccion, A. V. A. (1998). *Consumer sensory testing for product development*. Maryland: Aspen Publications.

Schwartz, N., & Pratt, C. H. (1956). Simultaneous vs successive presentation in a paired comparison situation. *Food Research*, *21*, 103–108.

Stone, H., & Sidel, J. L. (2004). *Sensory evaluation practices*, (3rd ed.). Oxford U.K.: Elsevier.

Sung, Y. E., Lee, H.-S., O'Mahony, M., & Kim, K.-O. (2011). Paired preference tests: Use of placebo stimuli with liking and buying preference. *Journal of Sensory Studies*, *26*, 106–117.

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. (2010). 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*, 60–65.

Xia, Y., Rivera-Quintero, A., Calderón, E., Zhong, F., & O'Mahony, M. (2014). Paired preference tests with reversed hidden demand characteristics. *Journal of Sensory Studies*, *29*, 149–158.

Xia, Y., Zhong, F., & O'Mahony, M. (2016). Paired preference testing: False preferences and disruptive protocols. *Food Science and Biotechnology*, *25*, 1–10.