Chapter 4

Discrimination Testing

Abstract Discrimination tests in most situations will only allow the sensory specialist to determine that two products perceptibly differ from one another or not. In this chapter we describe the more familiar discrimination tests such as paired comparison, duo–trio, triangle, dual standard, and A-not-A, as well as less used tests such as ABX and sorting tests. Data analysis techniques for these tests are described in detail (binomial, chi-square, *z*-, and beta-binomial distributions). Additionally, we begin the discussion of the effect of statistical power in sensory tests—this is further discussed in Chapter 5 and the Appendix of the book. The need for replication in sensory discrimination tests and the analysis of these data are discussed. Lastly, we discuss the need for warm-up samples in certain situations and well as some common issues arising from the interpretation of the results of sensory discrimination tests.

Chance favors only those who knows how to court her.
—Charles Nicolle

Contents

4.1 Discrimination Testing

Discrimination tests should be used when the sensory specialist wants to determine whether two samples are perceptibly different (Amerine et al., 1965; Meilgaard et al., 2006; Peryam 1958; Stone and Sidel, 2004). It is possible for two samples to be chemically different in formulation but for humans not to perceive this difference. Product developers exploit this possibility when they reformulate a product by using different ingredients while simultaneously not wanting the consumer to detect a difference. For example, an ice cream manufacturer may want to substitute the expensive vanilla flavor used in their premium vanilla ice cream with a cheaper vanilla flavor. However, they also may

H.T. Lawless, H. Heymann, *Sensory Evaluation of Food*, Food Science Text Series,
DOI 10.1007/978-1-4419-6488-5_4, © Springer Science+Business Media, LLC 2010

not want the consumer to perceive a difference in the product. A properly executed discrimination test with sufficient power indicating that the two ice cream formulations are not perceptibly different would allow the company to make the substitution with lowered risk. This is an ideal use of sensory discrimination testing. Discrimination testing may also be used when a processing change is made which the processor hopes would not affect the sensory characteristics of the product. In both of these cases the objective of the discrimination test is not to reject the null hypothesis, this is also known as similarity testing.

However, when a company reformulates a product to make a "new, improved" version then the discrimination test could be used to indicate that the two formulations are perceived to be different. In this case the objective of the discrimination is to reject the null hypothesis. If the data indicate that the two formulations are perceptibly different then the sensory scientist has to do a test that would indicate that the "new" formulation is perceived to be an improvement by the targeted consumer (see Chapters 13–15).

If the difference between the samples is very large and thus obvious, discrimination tests are not useful. If preliminary bench testing indicates that the two samples will be perceptibly different to all panelists then these discrimination tests should not be used. In such cases it may be useful to do scaling techniques to indicate the exact magnitude of the difference between the samples (see Chapter 7). In other words, discrimination testing is most useful when the differences between the samples are subtle. However, these subtle differences make the risk of Type II errors more likely (see later in this chapter and Appendix E).

Discrimination tests are usually performed when there are only two samples. It is possible to do multiple difference tests to compare more than two products but this is not efficient or statistically defensible. Usually ranking or scaling techniques will prove to be more effective (see Chapter 7).

There are a number of different discrimination tests available including triangle tests, duo-trio tests, paired comparison tests, *n*-alternative forced choice tests, tetrad tests (Frijters, 1984), polygonal and polyhedral tests (Basker, 1980). In Chapter 1, we briefly outlined the history associated with the triangle, duo–trio, and paired comparison tests. In the following section the more usual discrimination tests and their uses are described in more detail.

4.2 Types of Discrimination Tests

See Table 4.1 for a summary of the types of available discrimination tests and Table 4.2 for the process of doing a discrimination test.

4.2.1 Paired Comparison Tests

There are two analytical sensory forms of this test, namely the directional paired comparison (also known as the two-alternative forced choice) test and the difference paired comparison (also known as the simple difference or the same/different) test. The decision to use one or the other form is dependent on the objective of the study. If the sensory scientist knows that the two samples differ only in a specific sensory attribute then the two-alternative forced choice (2-AFC) method is used. In fact, as we will discuss in Chapter 5, it is always more efficient and powerful to use a directional paired comparison test specifying the sensory attribute in which the samples differ (if known) than to ask the panelists to identify the different sample. On the other hand, if the sensory scientist does not know in which sensory attribute(s) the samples differ than other techniques, such as the difference paired comparison must be employed, despite the subsequent loss of power.

For both paired comparison methods the probability of selection of a specific product, by chance alone (guessing), is one chance in two. However, as explained in Chapter 5 the situation is a little fuzzier for the same/different test where the probability is affected by the individual panelist's decision criterion. In both cases the null hypothesis states that in the long run (across all possible replications and samples of people) when the underlying population cannot discriminate between the samples they will pick each product an equal number of times. Thus the probability of the null hypothesis is $P_{\text{pc}} = 0.5$. Remember that P_{pc}, the proportion that we are making an inference (a conclusion) about, refers to the proportion we would see correct in the underlying population (and not the proportion correct in our sample or the data). That is why statistical hypothesis testing is part of inferential statistics. What the null hypothesis states in mathematical terms can also be verbalized as follows: If the

Table 4.1 Types of available discrimination tests

Class of test	Test	Samples:	Samples: test phase	Task/instructions	Chance
		inspection phase			probability
Oddity	Triangle	(None)	A, A', B (or A, B, B')	Choose the most different sample	1/3
Matching	Constant reference duo–trio	Ref-A	A, B	Match sample to reference	1/2
	Balanced reference duo–trio	Ref-A, Ref-B	A, B	Match sample to reference	1/2
	ABX	Ref-A, Ref-B	A (or B)	Match sample to reference	1/2
	Dual standard	Ref-A, Ref-B	A, B	Match both pairs	1/2
Forced choice	Paired comparison	(None)	A, B	Choose sample with most of specified attribute	1/2
	3-AFC	(None)	A, A',B	(Same)	1/3
	n-AFC	(None)	A_1-A_{n-1}, B	(Same)	1/*n*
	Dual pair	(None)	A, B and A, A'	Choose A, B (different pair)	1/2
Sorting	Two out of five	(None)	A, A', B, B', B''	Sort into two groups	1/10
	4/8 "Harris–Kalmus"	(None)	A_1-A_4, B_1-B_4	Sort into two groups	1/70
Yes/no	Same–different	(None)	Pairs: A, A' or A, B	Choose response: "Same" or "different"	N/A[a]
(Response choice)	A, not-A	Ref-A	A or B	Choose response: "A" or "not-A"	N/A[a]

[a] For the yes/no tests, a criterion may be set by each individual and therefore the probability may not be equal to 1/2. See Chapter 5 for further discussion of criterion in yes/no tasks

Table 4.2 Steps in conducting a difference test

1. Obtain samples and confirm test purpose, details, timetable, and panelists' training (e.g., training with the process) with client.
2. Decide testing conditions (sample size, volume, temperature, etc.) and clear with client.
3. Write instructions to the panelists and construct ballot.
4. Recruit potential panelists.
5. Screen panelists for acuity.
6. Train to do specific difference test (can use colors or shapes or spiked samples).
7. Set up counterbalanced orders.
8. Assign random three-digit codes and label sample cups/plates.
9. Conduct test.
10. Analyze results.
11. Communicate results to client or end user.

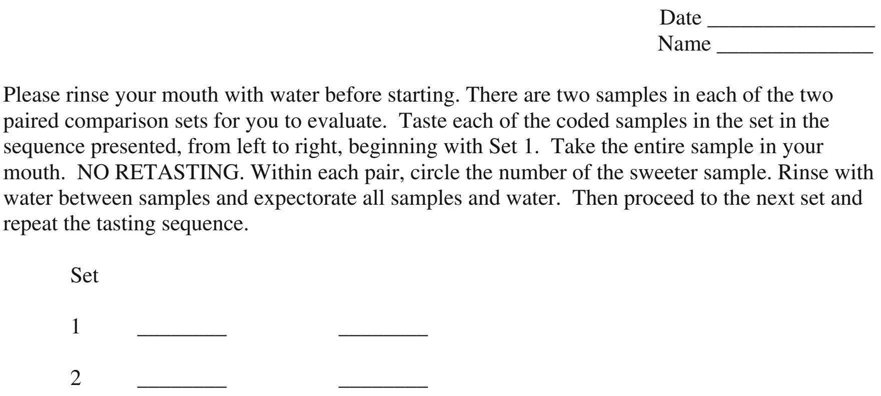
underlying population cannot discriminate between the samples then the probability of choosing sample A (that is the P_A) is equal to the probability of choosing sample B (P_B). Mathematically, this may be written as

$$H_0 : P_A = P_B = \frac{1}{2} \qquad (4.1)$$

However, as we will see the verbal forms of the alternate hypotheses for the two paired comparison tests differ.

4.2.1.1 Directional Paired Comparison Method (or the Two-Alternative Forced-Choice Method)

In this case, the experimenter wants to determine whether the two samples differ in a specified dimension, such as sweetness, yellowness, crispness. The two samples are presented to the panelist simultaneously and the panelist is asked to identify the sample that is higher in the specified sensory attribute. Figure 4.1 shows a sample score sheet. The panelist must clearly understand what the sensory specialist



Date ______________
Name ______________

Please rinse your mouth with water before starting. There are two samples in each of the two paired comparison sets for you to evaluate. Taste each of the coded samples in the set in the sequence presented, from left to right, beginning with Set 1. Take the entire sample in your mouth. NO RETASTING. Within each pair, circle the number of the sweeter sample. Rinse with water between samples and expectorate all samples and water. Then proceed to the next set and repeat the tasting sequence.

Set

1 ________ ________

2 ________ ________

Fig. 4.1 Example of a directional paired comparison (2-AFC) score sheet.

means by the specified dimension and the panelist should therefore be trained to identify the specified sensory attribute. The panelist should also be trained to perform the task as described by the score sheet. The directional paired comparison test has two possible serving sequences (AB, BA). These sequences should be randomized across panelists with an equal number of panelists receiving either sample A or sample B first.

The test is one tailed since the experimenter knows which sample is supposed to be higher in the specified dimension. The alternative hypothesis for the directional paired comparison test is that if the underlying population can discriminate between the samples based on the specified sensory attribute then the sample higher in the specified dimension (say A) will be chosen more often as higher in intensity of the specified dimension than the other sample (say B), this is P_{pc}. Mathematically this may be written as Eq. (4.2)

$$H_{\mathrm{A}} : P_{\mathrm{pc}} > \frac{1}{2} \tag{4.2}$$

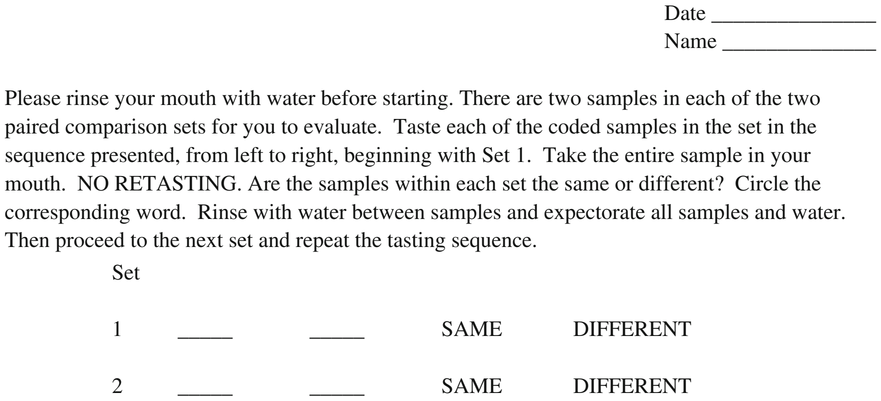
The results of the paired directional (2-AFC) test indicate the direction of the specified difference between the two samples. The sensory specialist must be sure that the two samples *only* differ in the single specified sensory dimension. This is often a problem with sensory discrimination testing of foods because changing one parameter frequently affects many other sensory attributes of the products. For example, removing some of the sugar from a sponge cake will likely make the cake less sweet but it would also affect the texture and the browning of the cake. In this case the directional paired comparison would not be an appropriate discrimination test to use.

4.2.1.2 Difference Paired Comparison (or the Simple Difference Test or the Same/Different Test)

This technique is similar to the triangle and duo–trio tests but it is not often used. It is best used, instead of the triangle or duo–trio test, when the product has a lingering effect or is in short supply and the presentation of three samples simultaneously would not be feasible (Meilgaard et al., 2006). In this case, the experimenter wants to determine whether the two samples differ without specifying the dimension(s) of the potential difference. An example would be if the study involves two sponge cakes, identical in formulation, except for the amount of sugar used. It is likely that the two cakes will differ in sweetness but probably also in texture and crust color.

The panelists are presented simultaneously with the two samples and are asked whether they perceive the samples to be the same or different. See Fig. 4.2 for a sample score sheet. The panelists only need to compare the two samples and decide whether they are similar or different. Humans easily make these types of comparisons and thus the task is relatively easy for the panelists. Thus, the panelists must be trained to understand the task as described by the score sheet but they need not be trained to evaluate specified sensory dimensions. The difference paired comparison method has four possible serving sequences (AA, BB, AB, BA). These sequences should be randomized across panelists with each sequence appearing an equal number of times.

The test is one tailed since the experimenter knows the correct answer to the question asked of each of the panelists, i.e., whether the two samples served to a



Date _______________
Name _______________

Please rinse your mouth with water before starting. There are two samples in each of the two paired comparison sets for you to evaluate. Taste each of the coded samples in the set in the sequence presented, from left to right, beginning with Set 1. Take the entire sample in your mouth. NO RETASTING. Are the samples within each set the same or different? Circle the corresponding word. Rinse with water between samples and expectorate all samples and water. Then proceed to the next set and repeat the tasting sequence.

Set

1 _____ _____ SAME DIFFERENT

2 _____ _____ SAME DIFFERENT

Fig. 4.2 Example of a difference paired comparison score sheet.

specific panelists were the same or different. The alternative hypothesis for the difference paired comparison test states that the samples are perceptibly different and that the population will correctly indicate that the samples are the same or different more frequently than 50% of the time. The mathematical form is

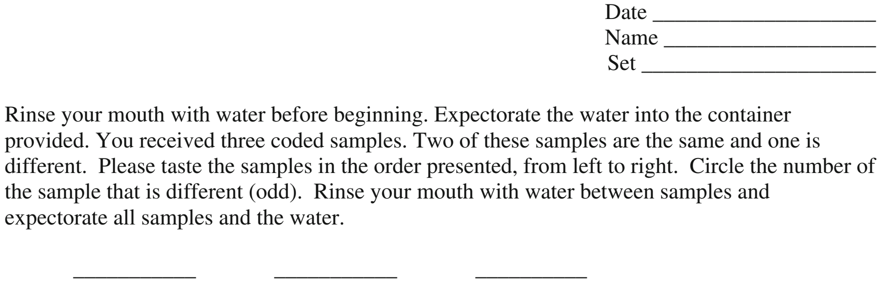
$$H_A : P_{pc} > \frac{1}{2} \tag{4.3}$$

The verbal form of the alternative hypothesis is that the population would be correct (saying that AB and BA pairs are different and that AA and BB pairs are the same) more than half the time. The results of the paired difference test will only indicate whether the panelists could significantly discriminate between the samples. Unlike the paired directional test, no specification or direction of difference is indicated. In other words, the sensory scientist will only know that the samples are perceptibly different but not in which attribute(s) the samples differed. An alternative analysis is presented in the Appendix to this chapter, where each panelist sees an identical pair (AA or BB) and one test pair (AB or BA) in randomized sequence.

4.2.2 Triangle Tests

In the triangle test, three samples are presented simultaneously to the panelists, two samples are from the same formulation and one is from the different formulation. Each panelist has to indicate either which sample is the odd sample or which two samples are most similar. The usual form of the score sheet asks the panelist to indicate the odd sample. However, some sensory specialists will ask the panelist to indicate the pair of similar samples. It probably does not matter which question is asked. However, the sensory specialist should not change the format when re-using panelists since they will get confused. See Fig. 4.3 for a sample score sheet. Similarly to the paired difference test the panelist must be trained to understand the task as described by the score sheet.

The null hypothesis for the triangle test states that the long-run probability (P_t) of making a correct selection when there is no perceptible difference between the samples is one in three (H_0:P_t = 1/3). The alternative hypothesis states that the probability that the underlying population will make the correct decision



Date ____________________
Name ___________________
Set _____________________

Rinse your mouth with water before beginning. Expectorate the water into the container provided. You received three coded samples. Two of these samples are the same and one is different. Please taste the samples in the order presented, from left to right. Circle the number of the sample that is different (odd). Rinse your mouth with water between samples and expectorate all samples and the water.

___________ ___________ __________

Fig. 4.3 Example of a triangle score sheet.

when they perceive a difference between the samples will be larger than one in three.

$$H_A : P_t > \frac{1}{3} \tag{4.4}$$

This is a one-sided alternative hypothesis and the test is one tailed. In this case there are six possible serving orders (AAB, ABA, BAA, BBA, BAB, ABB) which should be counterbalanced across all panelists. As with the difference paired comparison, the triangle test allows the sensory specialist to determine if two samples are perceptibly different but the direction of the difference is not indicated by the triangle test. Again, the sensory scientist will only know that the samples are perceptibly different but not in which attribute(s) the samples differed.

4.2.3 Duo–Trio Tests

In the duo–trio tests, the panelists also receive the three samples simultaneously. One sample is marked reference and this sample is the same formulation as one of the two coded samples. The panelists have to pick the coded sample that is most similar to reference. The null hypothesis states that the long-run probability (P_{dt}) of the population making a correct selection when there is no perceptible difference between the samples is one in two (H_0: $P_{dt} = 1/2$). The alternate hypothesis is that if there is a perceptible difference between the samples the population would match the reference and the sample correctly more frequently than one in two times.

$$H_A : P_{dt} > \frac{1}{2} \tag{4.5}$$

Again, the panelists should be trained to perform the task as described by the score sheet correctly. Duo–trio tests allow the sensory specialist to determine if two samples are perceptibly different but the direction of the difference is not indicated by the duo–trio test. In other words, the sensory scientist will only know that the samples are perceptibly different but not in which attribute(s) the samples differed.

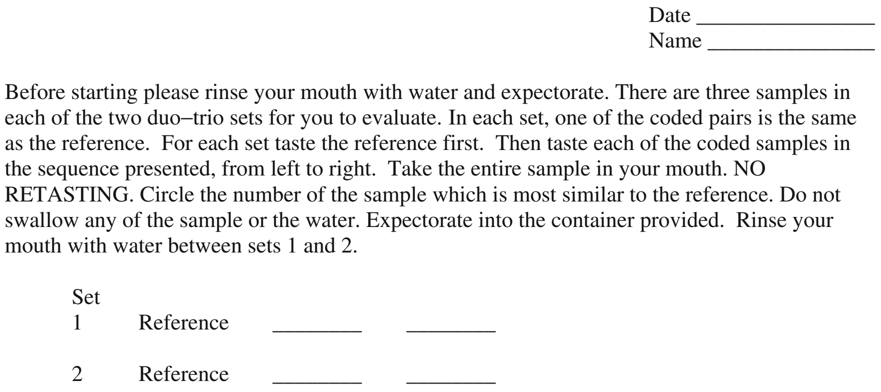
There are two formats to the duo–trio test, namely the constant reference duo–trio test and the balanced reference duo–trio test. From the point of view of the panelists the two formats of the duo–trio test are identical (see Figs. 4.4a and b), but to the sensory specialist the two formats differ in the sample(s) used as the reference.

4.2.3.1 Constant Reference Duo–Trio Test

In this case, all panelists receive the same sample formulation as the reference. The constant reference duo–trio test has two possible serving orders (R_A BA, R_A AB) which should be counterbalanced across all panelists. The constant reference duo–trio test seems to be more sensitive especially if the panelists have had prior experience with the product (Mitchell, 1956). For example, if product X is the current formulation (familiar to the panelists) and product Z is a new reformulation then a constant reference duo–trio test with product X as reference would be the method of choice.

4.2.3.2 Balanced Reference Duo–Trio Test

With the balanced reference duo–trio test half of the panelists receive one sample formulation as the



Date ________________
Name ________________

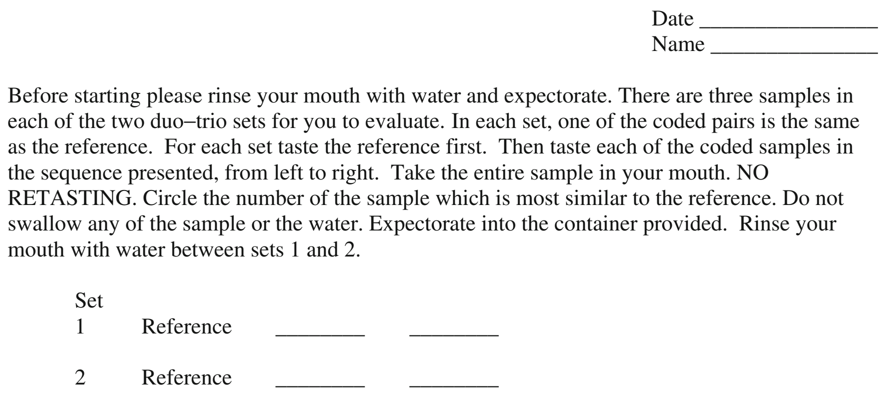
Before starting please rinse your mouth with water and expectorate. There are three samples in each of the two duo–trio sets for you to evaluate. In each set, one of the coded pairs is the same as the reference. For each set taste the reference first. Then taste each of the coded samples in the sequence presented, from left to right. Take the entire sample in your mouth. NO RETASTING. Circle the number of the sample which is most similar to the reference. Do not swallow any of the sample or the water. Expectorate into the container provided. Rinse your mouth with water between sets 1 and 2.

Set
1 Reference ________ ________

2 Reference ________ ________

Fig. 4.4a Example of a constant reference duo–trio score sheet.

Fig. 4.4b Example of a balanced reference duo–trio score sheet.



Date ________________
Name ________________

Before starting please rinse your mouth with water and expectorate. There are three samples in each of the two duo–trio sets for you to evaluate. In each set, one of the coded pairs is the same as the reference. For each set taste the reference first. Then taste each of the coded samples in the sequence presented, from left to right. Take the entire sample in your mouth. NO RETASTING. Circle the number of the sample which is most similar to the reference. Do not swallow any of the sample or the water. Expectorate into the container provided. Rinse your mouth with water between sets 1 and 2.

Set
1 Reference ________ ________

2 Reference ________ ________

reference and the other half of the panelists receive the other sample formulation as the reference. In this case, there are four possible serving orders (R_A BA, R_A AB, R_B AB, R_B BA) which should be counterbalanced across all panelists. This method is used when both products are prototypes (unfamiliar to the panelists) or when there is not a sufficient quantity of the more familiar product to perform a constant reference duo–trio test.

4.2.4 n-Alternative Forced Choice (n-AFC) Methods

The statistical advantages and hypotheses associated with and the uses of the *n*-AFC tests will be discussed in detail in Chapter 5. As we have seen the 2-AFC method is the familiar directional paired comparison method. The three-alternative forced choice (3-AFC) method is similar to a "directional" triangle method where the panelists receive three samples simultaneously and are asked to indicate the sample(s) that are higher or lower in a specified sensory dimension (Frijters, 1979). In any specific 3-AFC study there are only three possible serving orders (AAB, ABA, BAA or BBA, BAB, ABB) that should be counterbalanced across all panelists. As with the 2-AFC the specified sensory dimension must be the only perceptible dimension in which the two samples may differ. The panelists must be trained to identify the sensory dimension evaluated. They must also be trained to perform the task as described by the score sheet (Fig. 4.5).

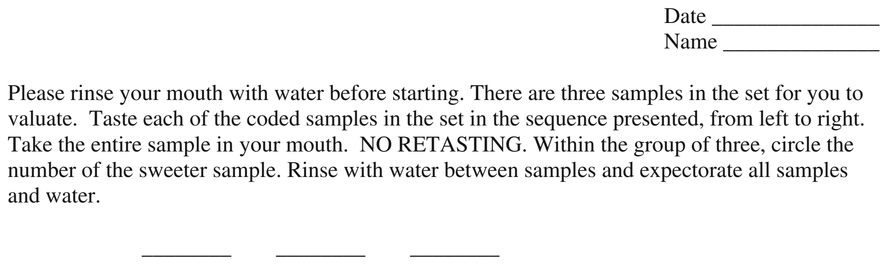
The three-alternative forced choice test will allow the sensory scientist to determine if the two samples differ in the specified dimension and which sample is higher in perceived intensity of the specified attribute. The danger is that other sensory changes will occur in a food when one attribute is modified and these may obscure the attribute in question. Another version of the *n*-AFC asks panelists to pick out the weakest or strongest in overall intensity, rather than in a specific attribute. This is a very difficult task for panelists when they are confronted with a complex food system.

4.2.5 A-Not-A tests

There are two types of A-not-A tests referenced in the literature. The first and the more commonly used version has a training phase with the two products followed by monadic evaluation phase (Bi and Ennis, 2001a, b), we will call this the standard A-not-A test. The second version is essentially a sequential paired difference test or simple difference test (Stone and Sidel, 2004), which we will call the alternate A-not-A test. The alternate A-not-A test is not frequently used. In the next section we will discuss the alternate A-not-A test first since the statistical analysis for this version is similar to that of the paired comparison discrimination test. The statistical analyses for the various standard A-not-A tests are based on a different theory and somewhat more complex and will be discussed later.

4.2.5.1 Alternate A-Not-A test

This is a sequential same/difference paired difference test where the panelist receives and evaluates the first



Date _______________
Name _______________

Please rinse your mouth with water before starting. There are three samples in the set for you to valuate. Taste each of the coded samples in the set in the sequence presented, from left to right. Take the entire sample in your mouth. NO RETASTING. Within the group of three, circle the number of the sweeter sample. Rinse with water between samples and expectorate all samples and water.

________ ________ ________

Fig. 4.5 Example of a three-alternative forced choice score sheet.

sample, that sample is then removed. Subsequently, the panelist receives and evaluates the second sample. The panelist is then asked to indicate whether the two samples were perceived to be the same or different. Since the panelists do not have the samples available simultaneously they must mentally compare the two samples and decide whether they are similar or different. Thus, the panelists must be trained to understand the task as described by the score sheet but they need not be trained to evaluate specified sensory dimensions. The alternate A-not-A test, like the difference paired comparison method, has four serving sequences (AA, BB, AB, BA). These sequences should be randomized across panelists with each sequence appearing an equal number of times. The test is one tailed since the experimenter knows the correct answer to the question asked of the panelists namely whether the two samples are the same or different. The null hypothesis of the alternate A-not-A test is the same as the difference paired comparison null hypothesis (H_0: $P_{\text{pc}} = 0.5$). The alternative hypothesis for this form of the A-not-A test is that if the samples are perceptibly different the population will correctly indicate that the samples are the same or different more frequently than one in two times. This alternative hypothesis is also the same as that of the difference paired comparison test (H_A: $P_{\text{pc}} > 1/2$).

The results of the A-not-A test only indicate whether the panelists could significantly discriminate between the samples when they are not presented simultaneously. Like the paired difference test, no direction of difference is indicated. In other words, the sensory scientist will only know that the samples are perceptibly different but not in which attribute(s) the samples differed.

This version of the A-not-A test is frequently used when the experimenter cannot make the two formulations have exactly the same color or shape or size, yet the color or shape or size of the samples are not relevant to the objective of the study. However, the differences in color or shape or size have to be very subtle and only obvious when the samples are presented simultaneously. If the differences are not subtle the panelists are likely to remember these and they will make their decision based on these extraneous differences.

4.2.5.2 Standard A-Not-A Test

Panelists inspect multiple examples of products that are labeled “A” and usually also products that are labeled “not-A.” Thus there is a learning period. Then once the training period has been completed the panelists receive samples one at a time and are asked whether each one is either A or not-A. As discussed by Bi and Ennis (2001a) the standard A-not-A test potentially has four different designs. For the monadic A-not-A test the panelist, after the training phase, is presented with a single sample (either A or not-A). In the paired A-not-A version the panelist, after completion of the training phase, is presented with a pair of samples, sequentially (one A and one not-A, counter balanced across panelists). In the replicated monadic A-not-A version the panelist, after completion of training, receives a series of samples of either A or not-A but not both. This version is rarely used in practice. Lastly, in the replicated mixed A-not-A version the panelist, after completion of training, receives a series of A and not-A samples. Each of these different formats requires different statistical models and using an inappropriate model could lead to a misleading conclusion. As described by Bi and Ennis (2001a) “The statistical models for the A-Not A method are different from that of other discrimination methods such as the *m*-AFC, the triangle, and the duo–trio methods.”

“Pearson’s and McNemar’s chi-square statistics with one degree of freedom can be used for the standard A-Not A method while binomial tests based on the proportion of correct responses can be used for the *m*-AFC, the triangle, and the duo–trio methods. The basic difference between the two types of difference tests is that the former involves a comparison of two proportions (i.e., the proportion of "A" responses for the A sample versus that for the Not A sample) or testing independence of two variables (sample and response) while the latter is a comparison of one proportion with a fixed value (i.e., the proportion of correct responses versus the guessing probability)". Articles by Bi and Ennis (2001a, b) clearly describe data analysis methods for these tests.. Additionally, the article by Brockhoff and Christensen (2009) describes a R-package called SensR (http://www.cran.r-project.org/package=sensR/) that may be used for the data analyses of some Standard A-not-A tests. The data analyses associated with the standard A-not-A tests are beyond the scope of this textbook, but see the Appendix of this chapter which shows the application of the McNemar chi-square for a simple A-not-A test where each panelist received one standard product (a "true" example of A) and one test product. Each is presented separately and a judgment is collected for both products.

4.2.6 Sorting Methods

In sorting tests the panelists are given a series of samples and they are asked to sort them into two groups. The sorting tests can be extremely fatiguing and are not frequently used for taste and aroma sensory evaluation but they are used when sensory specialists want to determine if two samples are perceptibly different in tactile or visual dimensions. The sorting tests are statistically very efficient since the long-run probability of the null hypotheses of the sorting tests can be very small. For example, the null hypothesis of the two-out-of-five test is 1 in 10 ($P_{2/5} = 0.1$) and for the Harris–Kalmus test the null hypothesis is 1 in 70 ($P_{4/8} = 0.0143$). These tests are discussed below.

4.2.6.1 The Two-Out-of-Five Test

The panelists receive five samples and are asked to sort the samples into two groups, one group should contain the two samples that are different from the other three samples (Amoore et al., 1968). Historically, this test was used for odor threshold work where the samples were very weak and therefore not very fatiguing (Amoore, 1979). The probability of correctly choosing the correct two samples from five by chance alone is equal to 0.1. This low probability of choosing the correct pair by chance is the main advantage of the method. However, major disadvantage of this method is the possibility of sensory fatigue. The panelists would have to make a number of repeat evaluations and this could be extremely fatiguing for samples that have to be smelled and tasted. This technique works well when the samples are compared visually or by tactile methods but it is usually not appropriate for samples that must be smelled or tasted. Recently Whiting et al. (2004) compared the two-out-of-five and the triangle test in determining perceptible differences in the color of liquid foundation cosmetics. They found that the triangle test results gave weak correlations with the instrumental color-differences but that the results of the two-out-of-five test were well correlated with the instrumental values.

4.2.6.2 The Harris–Kalmus Test

The Harris–Kalmus test was used to determine individual thresholds for phenyl thiocarbamide (PTC, a.k.a. phenyl thiourea, PTU). In this test panelists are exposed to increasing concentration levels of PTC in groups of eight (four samples containing water and four samples containing the current concentration of PTC). The panelists are asked to sort the eight samples into two groups of four. If the panelist does the sorting task incorrectly he/she is then exposed to the next higher concentration of PTC. The sorting task continues until the panelist correctly sorts the two groups of four samples. That concentration level of PTC is then identified as the threshold level for that panelist (Harris and Kalmus, 1949–1950). The method has the same disadvantage as the two-out-of-five test, in that it could be fatiguing. However, as soon as the panelist correctly sorts the samples the researcher concludes that the panelist is sensitive to PTC. Panelists insensitive to PTC only "taste" water in the solutions and are thus not fatigued. A shortened version of this test using three-out-of-six was used by Lawless (1980) for PTC and PROP (6-*n*-propyl thiouracil) thresholds.

4.2.7 The ABX Discrimination Task

The ABX discrimination task, as its name intends to suggest, is a matching-to-sample task. The panelist receives two samples, representing a control sample and a treatment sample. As in other discrimination tasks, the "treatment" in food research is generally an ingredient change, a processing change or a variable having to do with packaging or shelf life. The "X" sample represents a match to one of the two inspected samples and the panelist is asked to indicate which one is the correct match. The chance probability level is 50% and the test is one tailed, as the alternative hypothesis is performance in the population above 50% (but not below). In essence, this task is a duo–trio test in reverse (Huang and Lawless, 1998). Instead of having only one reference, two are given, as in the dual standard discrimination test. In theory, this allows the panelists to inspect the two samples and to discover for themselves the nature of the sensory difference between the samples, if any. As the differences are completely "demonstrated" to the panelists, the task should enjoy the same advantage as the dual standard test (O'Mahony et al., 1986) in that the participants should be able to focus on one or more attributes of difference and use these cues to match the test item to the correct sample. The inspection process of the two labeled samples may also function as a warm-up period. The test may also have some advantage over the dual standard test since only one item, rather than two are presented, thus inducing less sensory fatigue, adaptation, or carry-over effects. On the other hand, giving only one test sample provides less evidence as to the correct match, so it is unknown whether this test would be superior to the dual standard. As in other general tests of overall difference (triangle, duo–trio) the nature of the difference is not specified and this presents a challenge to the panelists to discover relevant dimensions of sensory difference and not be swayed by apparent but random differences. As foods are multi-dimensional, random variation in irrelevant dimensions can act as a false signal to the panelists and draw their attention to sensory features that are not consistent sources of difference (Ennis and Mullen, 1986).

This test has been widely used as a forced choice measure of discrimination in psychological studies, for example, in discrimination of speech sounds and in measuring auditory thresholds (Macmillan et al., 1977; Pierce and Gilbert, 1958). Several signal detection models (see Chapter 5) are available to predict performance using this test (Macmillan and Creelman, 1991). The method has been rarely if ever applied to food testing, although some sensory scientists have been aware of it (Frijters et al., 1980). Huang and Lawless (1998) did not see any advantages to the use of this test over more standard discrimination tests.

4.2.8 Dual-Standard Test

The dual standard was first used by Peryam and Swartz (1950) with odor samples. It is essentially a duo–trio test with two reference standards—the control and the variant. The two standards allow the panelists to create a more stable criterion as to the potential difference between the samples. The potential serving orders for this test are $R_{(A)}$ $R_{(B)}$, AB, $R_{(A)}$ $R_{(B)}$ BA, $R_{(B)}$ $R_{(A)}$ AB, $R_{(B)}$ $R_{(A)}$ BA. The probability of guessing the correct answer by chance is 0.5 and the data analyses for this test are identical to that of the duo–trio test. Peryam and Swartz felt quite strongly that the technique would work best with odor samples due to the relatively quick recovery and that the longer recovery associated with taste samples would preclude the use of the test. The test was used by Pangborn and Dunkley (1966) to detect additions of lactose, algin gum, milk salts, and proteins to milk. O'Mahony et al. (1986) working with lemonade found that the dual-standard test elicited superior performance over the duo–trio test. But O'Mahony (personal communication, 2009) feels that this result is in error, since the panelists were not instructed to evaluate the standards prior to each pair evaluation and therefore the panelists were probably reverting to a 2-AFC methodology. This would be in agreement with Huang and Lawless (1998) who studied sucrose additions to orange juice and they did not find superiority in performance between the dual standard and the duo–trio or the ABX tests.

4.3 Reputed Strengths and Weaknesses of Discrimination Tests

If the batch-to-batch variation within a sample formulation is as large as the variation between formulations then the sensory specialist should not use triangle or duo–trio tests (Gacula and Singh, 1984). In this case the paired comparison difference test could be used but the first question that the sensory specialist should ask is whether the batch-to-batch variation should not be studied and improved prior to any study of new or different formulations.

The major weakness of all discrimination tests is that they do not indicate the magnitude of the sensory difference(s) between the sample formulations. As the simple discrimination tests are aimed at a yes/no decision about the existence of a sensory difference, they are not designed to give information on the magnitude of a sensory difference, only whether one is likely to be perceived or not. The sensory specialist should not be tempted to conclude that a difference is large or small based on the significance level or the probability (p-value) from the statistical analysis. The significance and p-value depend in part upon the number of panelists in the test as well as the inherent difficulty of the particular type of discrimination test method. So these are no acceptable indices of the size of the perceivable difference. However, it is sensible that a comparison in which 95% of the judges answered correctly has a larger sensory difference between control and test samples than a comparison in which performance was only at 50% correct. This kind of reasoning works only if a sufficient number of judges were tested, the methods were the same, and all test conditions were constant. Methods for interval level scaling of sensory differences based on proportions of correct discriminations in forced choice tests are discussed further in Chapter 5 as Thurstonian scaling methods. These methods are indirect measures of small differences. They are also methodologically and mathematically complex and require certain assumptions to be met in order to be used effectively. Therefore we feel that the sensory specialist is wiser to base conclusions about the degree of difference between samples on scaled (direct) comparisons, rather than indirect estimates from choice performance in discrimination tests. However, there are alternative opinions in the sensory community and we suggest that interested parties read Lee and O'Mahony (2007).

With the exception of the 2-AFC and 3-AFC tests the other discrimination tests also do not indicate the nature of the sensory difference between the samples. The major strength of the discrimination tests is that the task that the panelists perform is quite simple and intuitively grasped by the panelists. However, it is frequently the very simplicity of these tests that lead to the generation of garbage data. Sensory specialists must be very aware of the power, replication, and counterbalancing issues associated with discrimination tests. These issues are discussed later in this chapter.

4.4 Data Analyses

The data from discrimination tests may be analyzed by any of the following statistical methods. The three data analyses are based on the binomial, chi-square, or normal distributions, respectively. All these analyses assume that the panelists were forced to make a choice. Thus they had to choose one sample or another and could not say that they did not know the answer. In other words, each panelist either made a correct or incorrect decision, but they all made a decision.

4.4.1 Binomial Distributions and Tables

The binomial distribution allows the sensory specialist to determine whether the result of the study was due to chance alone or whether the panelists actually perceived a difference between the samples. The following formula allows the sensory scientists to calculate the probability of success (of making a correct decision; p) or the probability of failure (of making an incorrect decision; q) using the following formula.

$$P(y) = \frac{n!}{y!(n-y)!} p^y p^{n-y} \tag{4.6}$$

where

n = total number of judgments
y = total number of correct judgments
p = probability of making the correct judgment by chance

In this formula, $n!$ describes the mathematical factorial function which is calculated as $n \times (n–1) \times (n–2) \ldots \times 2 \times 1$. Before the widespread availability of calculators and computers, calculation of the binomial formula was quite complicated, and even now it remains somewhat tedious. Roessler et al. (1978) published a series of tables that use the binomial formula to calculate the number of correct judgments and their probability of occurrence. These tables make it very easy to determine if a statistical difference were detected between two samples in discrimination tests. However, the sensory scientist may not have these tables easily available thus he/she should also know how to analyze discrimination data using statistical tables that are more readily available. We abridged the tables from Roessler et al. (1978) into Table 4.3. Using this table is very simple. For example, in a duo–trio test using 45 panelists, 21 panelists correctly matched the sample to the reference. In Table 4.3, in the section for duo–trio tests, we find that the table value for 45 panelists at 5% probability is 29. This value is larger than 21 and therefore the panelists could not detect a difference between the samples. In a different study, using a triangle test, 21 of 45 panelists correctly identified the odd sample. In Table 4.3, in the section for triangle tests, we find that the table value for 45 panelists at 5% probability is 21. This value is equal to 21 and therefore the panelists could detect a significant difference between the samples at the alpha probability of 5%.

4.4.2 The Adjusted Chi-Square (χ^2) Test

The chi-square distribution allows the sensory scientist to compare a set of observed frequencies with a matching set of expected (hypothesized) frequencies. The chi-square statistic can be calculated from the following formula (Amerine and Roessler, 1983), which includes the number –0.5 as a continuity correction. The continuity correction is needed because the χ^2 distribution is continuous and the observed frequencies from discrimination tests are integers. It is not possible for one-half of a person to get the right answer and so the statistical approximation can be off by as much as ½, maximally.

$$\chi^2 = \left[\frac{(|O_2 - E_2| - 0.5)^2}{E_1}\right] + \left[\frac{(|O_2 - E_2| - 0.5)^2}{E_2}\right] \tag{4.7}$$

where

O_1 = observed number of correct choices
O_2 = observed number incorrect choices
E_1 = expected number of correct choices
E_1 is equal to total number of observations (n) times probability (p) of a correct choice, by chance alone in a single judgment where
$p = 0.100$ for the two-out-of-five test
$p = 0.500$ for duo–trio, paired difference, paired directional, alternate A-not-A tests
$p = 0.333$ for triangle tests
E_2 = expected number of incorrect choices
E_2 is equal to total number of observations (n) times probability (q) of an incorrect choice, by chance alone in a single judgment where $q = 1 - p$
$q = 0.900$ for the two-out-of-five test
$q = 0.500$ for duo–trio, paired difference, paired directional, alternate A-not-A tests, ABX tests
$q = 0.667$ for triangle tests

The use of discrimination tests allows the sensory scientist to determine whether two products are statistically perceived to be different, therefore the degrees of freedom equal one (1). Therefore, a χ^2 table using df = 1 should be consulted, for alpha (α) at 5% the critical χ^2 value is 3.84. For other alpha levels consult the chi-square table in the Appendix.

4.4.3 The Normal Distribution and the Z-Test on Proportion

The sensory specialist can also use the areas under the normal probability curve to estimate the probability of chance in the results of discrimination tests. The tables associated with the normal curve specify areas under the curve (probabilities) associated with specified values of the normal deviate (z). The following two formulae (Eqs. (4.8) and (4.9)) can be used to calculate the z-value associated with the results of a specific discrimination test (Stone and Sidel, 1978):

$$z = \frac{[P_{obs} - P_{chance}] - \frac{1}{2N}}{\sqrt{pq/N}} \tag{4.8}$$

where

$P_{obs} = X/N$
P_{chance} = probability of correct decision by chance
For triangle test: $P_{chance} = 1/3$

Table 4.3 Minimum numbers of correct judgments[a] to establish significance at probability levels of 5 and 1% for paired difference and duo–trio tests (one tailed, $p = 1/2$) and the triangle test (one tailed, $p = 1/3$)

Paired difference and duo–trio tests			Triangle test		
Number of trials (n)	Probability levels		Number of trials (n)	Probability levels	
	0.05	0.01		0.05	0.01
5	5	–	3	3	–
6	6	–	4	4	–
7	7	7	5	4	5
8	7	8	6	5	6
9	8	9	7	5	6
10	9	10	8	6	7
11	9	10	9	6	7
12	10	11	10	7	8
13	10	12	11	7	8
14	11	12	12	8	9
15	12	13	13	8	9
16	12	14	14	9	10
17	13	14	15	9	10
18	13	15	16	9	11
19	14	15	17	10	11
20	15	16	18	10	12
21	15	17	19	11	12
22	16	17	20	11	13
23	16	18	21	12	13
24	17	19	22	12	14
25	18	19	23	12	14
26	18	20	24	13	15
27	19	20	25	13	15
28	19	21	26	14	15
29	20	22	27	14	16
30	20	22	28	15	16
31	21	23	29	15	17
32	22	24	30	15	17
33	22	24	31	16	18
34	23	25	32	16	18
35	23	25	33	17	18
36	24	26	34	17	19
37	24	26	35	17	19
38	25	27	36	18	20
39	26	28	37	18	20
40	26	28	38	19	21
41	27	29	39	19	21
42	27	29	40	19	21
43	28	30	41	20	22
44	28	31	42	20	22
45	29	31	43	20	23
46	30	32	44	21	23
47	30	32	45	21	24
48	31	33	46	22	24
49	31	34	47	22	24
50	32	34	48	22	25
60	37	40	49	23	25
70	43	46	50	23	26

Table 4.3 (continued)

Paired difference and duo–trio tests			Triangle test		
Number of trials (n)	Probability levels		Number of trials (n)	Probability levels	
80	48	51	60	27	30
90	54	57	70	31	34
100	59	63	80	35	38
110	65	68	90	38	42
120	70	74	100	42	45
130	75	79	110	46	49
140	81	85	120	50	53
150	86	90	130	53	57
160	91	96	140	57	61
170	97	101	150	61	65
180	102	107	160	64	68
190	107	112	170	68	72
200	113	117	180	71	76
			190	75	80
			200	79	83

[a] Created in EXCEL 2007 using B. T. Carr's Discrimination Test Analysis Tool EXCEL program (used with permission)

For duo–trio and paired comparison tests: $P_{\text{chance}} = ½$
X = number of correct judgments
N = total number of judgments.

Alternately one can use the following equation:

$$z = \frac{X - np - 0.5}{\sqrt{npq}} \tag{4.9}$$

where

X = number of correct responses
n = total number of responses
p = probability of correct decision by chance
For triangle test: $p = 1/3$
For duo–trio and paired comparison tests: $p = ½$ and in both cases $q = 1-p$

As with the χ^2 calculation a continuity correction of –0.5 has to be made. Consult a Z-table (area-under-normal-probability curve) to determine the probability of this choice being made by chance. The critical Z-value for a one-tailed test at alpha (α) of 5% is 1.645. See the Z-table in the Appendix for other values.

4.5 Issues

4.5.1 The Power of the Statistical Test

Statistically, there are two types of errors that the sensory scientist of any sensory method can make when testing the null hypothesis (H_0). These are the Type I (α or alpha) and Type II (β or beta) errors (see Appendix E for a more extensive discussion). A Type I error occurs when the sensory scientist rejects the null hypothesis (H_0) when it is actually true. When making a Type I error in a discrimination test we would conclude that the two products are perceived to be different when they are actually not perceptibly different. The Type I error is controlled by the sensory scientist choice of the size of alpha. Traditionally, alpha is chosen to be very low (0.05, 0.01, or 0.001) which means that there is a 1 in 20, 1 in 100, and 1 in 1,000 chance, respectively, of making a Type I error. A Type II error occurs when the sensory scientist accepts the null hypothesis (H_0) when it is actually false. The Type II error is based on the size of and it is the risk of not finding a difference when one actually exists and it is defined as 1–beta. In other words, the power of a test could be defined as the probability of finding a difference if one actually exists or it is the probability of making the correct decision that the two samples are perceptibly different.

The power of the test is dependent on the magnitude of the difference between the samples, the size of alpha, and the number of judges performing the test.

4.5.1.1 Why Is Power Important When Performing Discrimination Tests?

A candy manufacturer wants to show that the new formulation of their peanut brittle is crunchier than the old formulation. Prior to the study the sensory scientist had decided which probability of making a Type I error (alpha) would be acceptable. If the sensory scientist had decided that alpha should be 0.01, then she/he had a 1 in 100 chance of committing a Type I error. Consider than that this candy maker performs a two-alternative forced choice test and the data indicate that the null hypothesis should be rejected. The sensory scientist is confronted with two possibilities. In the first case, the null hypothesis is actually false and should be rejected; in this case the new formulation is actually crunchier than the old formulation. In the second case, the null hypothesis is actually true and the sensory scientist has made a Type I error. In this type of study the Type I error is usually minimized because the sensory scientist wants to be quite certain that the new formulation is different from the old formulation. In this case the power of the test is only of passing interest.

Consider a second scenario. An ice cream manufacturer wants to substitute the expensive vanilla flavor used in their premium vanilla ice cream with a cheaper vanilla flavor. However, they do not want the consumer to perceive a difference in the product. They perform a triangle test to determine if a panel could tell a difference between the samples. The data indicate that the null hypothesis should not be rejected. Again the sensory scientist is confronted with two possibilities. In the first case, the null hypothesis is true and the two formulations are not perceptibly different. In the second case the samples are perceptibly different but the sensory scientist was making a Type II error. In this type of study the Type II error should be minimized (power of the test should be maximized) so that the sensory scientist can state with some confidence that the samples are not perceptibly different.

In many published discrimination studies the authors claim that a discrimination test indicated that two samples were not significantly different. Frequently, the power of these tests is not reported, but it can often be calculated post hoc. It is unfortunately the case that the power of these tests is often very low, suggesting that the research would not have revealed a difference even if a difference existed. Or in statistical jargon, the probability of a Type II error was high.

4.5.1.2 Power Calculations

Discrimination test power calculations are not simple. However, this does not absolve the sensory scientist from attempting to determine the power associated with a specific study, especially when the objective of the study is to make an equivalent ingredient substitution and therefore the objective of the study is not to reject the null hypothesis. In general, the sensory specialist should consider using a large sample size when power needs to be high ($N = 50$ or greater). This is essential in any situation where there are serious consequences to missing a true difference.

The sensory scientist will frequently make a post hoc power calculation. In this calculation the power of the test is calculated after the completion of the study. The sensory scientist can also make an a priori calculation of the number of judgments needed for a specific power. In both cases the sensory scientist must make a series of assumptions and all power calculations will only be as good as these assumptions. The scientists must be as extremely careful when making the required assumptions for the power calculations. A number of authors (Amerine et al., 1965; Ennis, 1993; Gacula and Singh, 1984; Kraemer and Thiemann, 1987; Macrae, 1995; Schlich, 1993) have studied power calculations for discrimination tests and have prepared tables that can be used to determine either the power of a specified discrimination test or the number of judgments needed for a specific level of power. The different tables (Ennis, 1993; Schlich, 1993) would lead one to slightly different conclusions as to the power of the test. The reason for these differences is that these calculations are based on a series of assumptions and slight differences in assumptions can lead to differences in power calculations. Power calculations will be discussed in more detail in Chapter 5 and Appendix E. Additionally the R-package SensR (http://www.cran.r-project.org/package=sensR/) written by Brockhoff and Christensen (2009) allows one to calculate the power associated with most discrimination tests.

4.5.2 Replications

As seen in the previous section and in the power section of Appendix E the number of judgments made in a discrimination test is very important. The number of judgments can be increased by using more panelists or by having a smaller number of panelists perform more tests. These two methods of increasing the number of judgments are clearly not equivalent. The ideal way to increase the number of judgments is to use more panelists. This is the only way in which the sensory specialist can be assured that all judgments were made independently. All the data analysis methods discussed above assume that all judgments were made entirely independently of one another (Roessler et al., 1978). Frequently and perhaps unfortunately, the industrial sensory scientist has only a limited number of panelists available. In this case the number of judgments may be increased by having each panelist evaluate the samples more than once in a session. In practice this is rather simply done. The panelist receives a set of samples which he/she evaluates. The samples and the score sheet are returned and then the panelist would receive a second set of samples. In some cases the panelist may even receive additional replications. It should be remembered that if the same panelists repeat their judgments on the same samples, there is a possibility these judgments are not totally independent. In other words, the replicate evaluations made by a specific individual may be related to each other. The use of replication increases the power of the difference test by increasing the number of judgments; however, depending on the assumptions relating to effect size (see Appendix E) and the type of difference test used the increase in power may be similar or less than when one uses the same number of independent judgments (Brockhoff, 2003). As can be seen in Table 4.4, extracted from Tables 3 and 4 in Brockhoff (2003) assuming an alpha of 5% and a medium effect size (37.5% above chance discriminating) for a triangle test the power for the independent judgments is more than for the replicated judgments. On the other hand for an alpha of 5% and a small effect size (25% above chance discriminator) for a duo–trio test the values are quite similar. Therefore replication of discrimination tests and the effect of this on power are not simple.

4.5.2.1 Analyzing Replicate Discrimination Tests

The important caution is that sensory scientists should not simply combine the data from replications and count the number of observations as the number of replicates multiplied by the number of panelists. They are not independent judgments and the situation needs to be examined closely before any such combination can be justified. There are a few simpler options that are available and a more complex option, the beta-binomial model.

Table 4.4 Limits of power (%) based on Monte Carlo simulations (extracted from Tables 3 and 4 of Brockhoff, 2003)

n^a	$k^b=1$	$k^b=2$	$k^b=3$	$k^b=4$	$k^b=5$
(a) Triangle test with alpha=5% and medium-effect size (37.5% above chance discriminator)					
12	40	70	81	90	91
24	74				
36	88				
48	97				
(b) Duo–trio test with alpha=5% and small-effect size (25% above chance discriminator)					
12	11	28	39	46	58
24	27				
36	37				
48	44				

[a]Number of panelists; [b]Number of replications

Simpler Options

First, the replications can be analyzed separately as independent tests. This can provide information as to whether there is a practice, warm-up, or learning effect if the proportion correct increases over trials. This could be useful information because a consumer in the real world will have multiple opportunities to interact with a product, and usually not just a single tasting. If later replications are statistically significant (and the first is not), that is usually grounds for concluding that the samples are in fact perceivably different. It is also possible, of course, that fatigue or adaptation or carry-over could have an effect on later replications, so the sensory scientist needs to consider the nature of each specific product and situation and make a reasoned judgment. If the replications lead to different results, further investigation or analysis may be necessary.

A second approach for duplicate tests is to simply tabulate the proportion of panelists that got both tests correct. Now the chance probability for a three sample test is 1/9 and for a test like the duo–trio or paired comparison, it becomes ¼. The same z-score formula applies for the binomial test on proportions (Eq. (4.10)) as

$$z = \frac{(P_{\text{obs}} - p) - 1/2n}{\sqrt{pq/n}} = \frac{(X - np) - 1/2}{\sqrt{npq}} \qquad (4.10)$$

where P_{obs} is the proportion correct ($=X/n$), X is the actual number correct, n is the number of judges, p is the chance probability, and $q = 1-p$.

Solving for X as a function of n, and using $p = 1/9$ for the triangle or 3-AFC tests, we get the following for $Z = 1.645$ ($p < 0.05$, one tailed):

$$X = n/9 + 0.517\sqrt{n} + 0.5 \qquad (4.11)$$

and for $p = ¼$ for the duplicated paired or duo–trio tests:

$$X = n/4 + 0.712\sqrt{n} + 0.5 \qquad (4.12)$$

These can easily be programmed on a spreadsheet, but do not forget to change the value of z if you wish to calculate the critical value of X for other probability levels. Of course, you must round up the value of X to the next highest integer since you are counting individuals. This approach is somewhat conservative, as it only considers those people who answered correctly on both tests, and it is possible that a person might miss the first test but get the replicate correct as a true discrimination. The solution to this issue (considering that some people might have partially correct discriminations) is to use a chi-square test to compare the observed frequencies against what one would expect by chance for zero, one or two correct judgments (e.g., 4/9, 4/9, and 1/9 for the replicated triangle or 3/8, 3/8, and ¼ for the replicated duo–trio).

4.5.2.2 Are Replications Statistically Independent?

Another approach to replication is to test whether the two replicates are in some way independent, seem to be varying randomly or whether there are systematic patterns in the data. One such approach is the test of Smith (1981), which can be used for two replications (i.e., a duplicate test). This test essentially determines whether there are significantly more correct choices on one replication, or whether they are not significantly different. It uses a binomial test on proportions with $p = ½$, so the binomial tables for the duo–trio test are applicable or for triangles it uses p=1/3, and thus the binomial tables for the triangle test would be applicable.

The total number of correct responses in each replication ($C1$, $C2$) are added together ($M = C1 + C2$). M represents the total number of trials (n) and either $C1$ or $C2$ (whichever is larger) is used to represent the number of correct responses in the study. If $C1$ or $C2$ is larger than the minimum number of judgments required for significance then the difference in proportions of correct responses between the two replications is significant and the replication data cannot be pooled. Each replication must then be analyzed independently. If the larger of $C1$ and $C2$ is less than the minimum number of judgments required for significance then the difference in proportions of correct responses between the two replications is not significant and the replication data can be pooled. The combined data can then be analyzed as if each judgment was made by a different panelist.

For example, a sensory specialist was asked to determine if two chocolate chip cookie formulations (one with sucrose as the sweetener and the other with a non-caloric sweetener) were perceptibly different from each other. The sensory specialist decided to use a constant reference duo−trio test to determine if the two formulations differed. A panel of 35 judges did the study in duplicate. In the first replication 28 judges correctly matched the reference and in the second replication 20 judges correctly matched the reference. The sensory scientist has to determine if the data can be pooled and if there is a significant perceived difference between the two cookie formulations.

Using Smith's test he found that $M = 28 + 20 = 48$, and that $C1 = 28$ is the larger of $C1$ and $C2$. Table 4.1 for duo–trio tests indicated that for $n = 48$ the minimum number of correct judgments for an alpha (α) of 5% is 31. Thus, 28 is less than 31 and the data from the two replications were pooled. The combined data were therefore 48 correct decisions out of a potential 70 (2×35). The sensory scientist decided to use the z-calculation to determine the exact probability of finding this result. Using Eq. (4.2), with 48 as the number of correct responses, with 70 as the total number of responses, with p equal to 1/2, $z = 2.9881$. The z-table showed that the exact probability of this value occurring by chance was 0.0028. The panelists could therefore perceive a difference between the two formulations.

The Beta-Binomial Model

The test devised by Smith does not address the issue of whether some panelists have different patterns across replicates than others. If people had systematic trends (e.g., some people getting both correct, easily discriminating and others missing consistently) you could still get a non-significant result by Smith's test, yet the data would hardly be independent from trial to trial. This issue is addressed in the beta-binomial model, which looks at patterns of consistency (versus. randomness) among the panelists. Although Smith's test is appropriate when detecting differences between replicates, it is not an airtight proof that replications are independent should the test not meet the significance level. The beta-binomial model allows us to pool replicates, but makes some adjustments in the binomial calculations to make the criteria more conservative when the data are not fully independent.

The beta-binomial model assumes that the performance of panelists is distributed like a beta distribution. This distribution has two parameters, but they can be summarized in a simple statistic called gamma. Gamma, which varies from zero to one, is a measure of the degree to which there are systematic behaviors of panelists (like always answering correctly or incorrectly), in which case gamma approaches one, or whether people seem to behave independently from trial to trial (gamma approaches zero). You can think of this as a kind of test of the independence of the observations, but from the perspective of individual performance, rather than group data as in Smith's test. Gamma is basically an estimate of the degree to which people's total number of correct choices varies from the panel mean. It is given by the following formula (Bi, 2006, p. 110):

$$\gamma = \frac{1}{r-1}\left\lfloor \frac{rS}{\mu(1-\mu)n} - 1 \right\rfloor \tag{4.13}$$

where

$r =$ the number of replicates
$S=$ a measure of dispersion
$\mu=$ mean proportion correct for the group (looking at each person's individual proportions as shown below)
$n =$ the number of judges. S and μ are defined as follows:

$$\mu = \frac{\sum_{i=1}^{n} x_i/r}{n} \tag{4.14}$$

where $x_i=$ the number of correct judgments summed across replicates for that panelist.

So µ is the mean of the number of correct replicates. S is defined as

$$S = \sum_{i=1}^{n} ((x_i/r) - \mu)^2 \tag{4.15}$$

Once gamma is found we have two choices. We can test to see whether the beta-binomial fits better than the simple binomial. This is essentially testing whether gamma is different from zero. The alternative is to go directly to beta-binomial tables for different levels of gamma. In these tables (see Table O), we combine replicates to get the total number of correct judgments, and compare that to the critical number required, for the total number of judgments (number of panelists times number of replicates). The tables adjust the binomial model requirements to be more conservative as gamma increases (i.e., as the panelists look less random and more systematic).

To test whether the beta-binomial is a better fit, we use the following Z-test (Bi, 2006, p. 114):

$$Z = \frac{E - nr}{\sqrt{2nr(r-1)}} \tag{4.16}$$

where E is another measure of dispersion defined as

$$E = \sum_{i=1}^{n} \frac{(x_1 - rm)^2}{m(1-m)} \tag{4.17}$$

and m is the mean proportion correct defined as

$$m = \sum_{i=1}^{n} x_i / nr \tag{4.18}$$

The advantage of doing the Z-test is that should you find a significant Z, then there is evidence that the panelists are not random, but that there are probably groups of consistent discriminators and also perhaps some people who are consistently not discriminating. In other words, a non-zero gamma is evidence for a consistently perceived difference for at least some of the panel! If the Z-test is non-significant, then one option is to pool replicates and just use the simple binomial table. Note that now you have effectively increased the sample size and given more power to the test. A good example of this approach can be found in Liggett and Delwiche (2005).

4.5.3 Warm-Up Effects

Much has been written concerning the potential advantages of giving warm-up samples before discrimination tests (e.g., Barbary et al., 1993; Mata-Garcia et al., 2007). A warm-up strategy involves repeated alternate tasting of the two different versions of the product, with the panelist's knowledge that they are different. Often (but not always) they are encouraged to try to figure out the way or ways in which the products differ. This is similar to giving a familiarization sample or dummy sample before the actual test, but with "warm up" it involves a more extended period of tasting. A single warm-up sample was in fact part of the original version of the duo–trio test (Peryam and Swartz, 1950).

However, evidence for the advantage of this added procedure is not very strong. In two early reports, a larger number of significant triangle tests were seen with warm-up for a wine sample and a fruit beverage (O'Mahony and Goldstein, 1986) and for NaCl solutions and for orange juices (O'Mahony et al., 1988). In the latter study it was unclear whether naming the difference gave any additional advantage. Later studies showed mixed results.

Thieme and O'Mahony (1990) found good discrimination for A, not-A, and paired comparison tests with warm-up but a direct comparison of the same kind of test, with and without warm-up, was lacking, so it is difficult to draw conclusions from that study. Dacremont et al. (2000) showed no effect of warm-up for the first trial of repeated triangles with naïve assessors nor with highly experienced judges. Panelists who were of intermediate experience did show some benefit of the warm-up. Kim et al. (2006) reported increased discrimination for the triangle, duo–trio, and same–different tests, but a 2-AFC test in which panelists were told to identify the NaCl sample (versus water) was also conducted before the warmed-up tests, so the cause of the increased discrimination in this study is not clear. Angulo et al. (2007) reported a small but non-significant increase in discriminability with a relatively less sensitive group in a 2-AFC test. Rousseau et al. (1999) looked at effects of a primer (single example) food and a familiarization with mustard samples before discrimination tests. The primer had no effect and the familiarization appeared to cause small increases in discriminability.

Taken together, these studies suggest that a warm-up protocol might have some benefits. The sensory practitioner should weigh the possible benefits against the extra burden on the panelist if the kind of extensive warm-up (three to ten pairs) that are usually done in the laboratory studies is adopted.

4.5.4 Common Mistakes Made in the Interpretation of Discrimination Tests

If a discrimination test had been performed properly, with adequate power and the sensory scientist finds that the two samples were not perceptibly different then there is no point in performing a subsequent consumer preference test with these samples. Logically, if two samples are sensorially perceived to be the same then one sample cannot be preferred over another. However, if a subsequent consumer preference test indicates that the one of the samples is preferred over the other, then the scientist must carefully review the adequacy of the discrimination test, especially the power of the test. Of course, any test is a sampling experiment, so there is always some probability of a wrong decision. Therefore a discrimination test does not “prove” that there is no perceivable difference and follow-up preference tests will sometimes be significant.

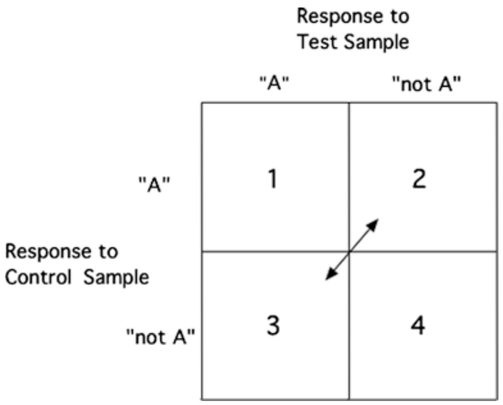
Sometimes, novice sensory scientists will do a preference tests and find that there was no significant preference for one sample over the other. This means that the two samples are liked or disliked equally. However, it does not mean that the samples are not different from one another. It is very possible for two samples to be perceptibly very different and yet to be preferred equally. For example, American consumers may prefer apples and bananas equally, yet that does not mean that apples are indistinguishable from bananas.

Appendix: A Simple Approach to Handling the A, Not-A, and Same/Different Tests

Both of these tests have response choices, rather than a sample choice. The choice of either response is affected by the criterion used by each panelist. For example, as a panelist one may ask oneself: Do I want to be really strict and be sure these products are different, or can I call them different if I think there is just an inkling of a difference? These criteria are clearly quite different from one another and will dramatically affect the outcome of the test. However, the sensory scientist does not know (not can he find out) which criterion each panelist used (sometimes even the panelists do not know since they do not explicitly decided on a criterion).

In order to get around this problem, we can give a control sample of true A in the A, not-A test or an identical pair (“same”) in the same/different test. The question then becomes whether the percent of choice of “not-A” for the test sample was greater than the choice of “not-A” for the control (i.e., true A) sample. Similarly, we can ask if the proportion of “different” responses was higher for the test pair than it was for the identical control pair. So we are comparing against a sensible baseline.

So far so good. A simple binomial test on proportions or a simple chi-square would seem to do it. But in most situations, we give both the true A and the test sample to the same person. In the same/different test, we would have given both a control pair (identical samples) and the test pair (samples that are physically different and might in fact be called “different”). The binomial test and the chi-square assume independent observations, but now we have two measurements on the same person (clearly NOT independent). So the appropriate statistic is provided by the McNemar test. Let us look at the A, not-A situation. We cast the data in a two-way table as follows, with everyone counted into one of the four cells (1, 2, 3, and 4 are the actual frequency counts, not percents):



	Response to Test Sample	
Response to Control Sample	"A"	"not A"
"A"	1	2
"not A"	3	4

Now, the people who are giving the same answer on both trials are not very interesting. They are not transmitting any information as to whether the products are different. They are counted in cells #1 and #4 above. The critical question is whether there are significantly more people in cell #2 (who call the test sample “not A” and the control sample “A”) than there are people who do the reverse in cell #3. If the cells have about the same counts, then let us face it, there is not much evidence for a difference. But if a lot of people call the test sample “not-A” and they recognize the control sample as a good example of what A should be like, then we have evidence that something important has in fact changed. The difference is perceivable.

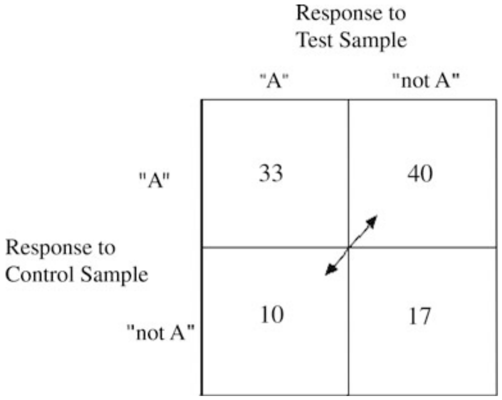
So we need to compare the size of cell 2 to cell 3. The McNemar test does just this for us. Let C_2 be the count in cell #2, and C_3 be the count in cell #3. Here is the formula:

$$\chi^2 \frac{(|C_2 - C_3| - 1)^2}{C_2 + C_3}$$

This χ^2 test has one degree of freedom, so the critical chi-square value that must be exceeded for significance is 3.84, if we use the standard alpha at 5%.

A worked example:

During an A-not-A test a group of 50 panelists were received (in randomized order: a control sample (the true A) and a test sample). The results are displayed in the figure below:



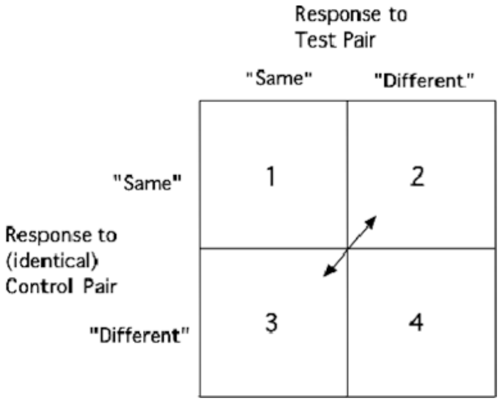
	"A"	"not A"
"A"	33	40
"not A"	10	17

$$\chi^2 \frac{(|40 - 10| - 1)^2}{40 + 10} = 16.82 \text{ which is larger than 3.84.}$$

The panelist therefore found a significant difference between the control and the test samples.

The same kind of chart can be drawn for the same–different test and the same comparison can be made:

If there are just a few more people in cell 3 than cell 2, it is probably random variation and there is no difference. If there is a LOT more people in cell 3 and you get a significant chi-square but in the "wrong direction," there is something wrong with your study (maybe you switched the codes, for example). Also, if there are a lot of people in cells 1 and 4, that is a concern because those folks are not distinguishing much, or maybe they have some "lazy" or lax criteria.



	"Same"	"Different"
"Same"	1	2
"Different"	3	4

References

Amerine, M. A., Pangborn, R. M. and Roessler, E. B. 1965. Principles of sensory evaluation. Academic, New York, NY.

Amerine, M. A. and Roessler, E. B. 1983. Wines, their sensory evaluation, Second Edition. W.H. Freeman, San Francisco, CA.

Amoore, J. E., Venstrom, D. and Davis, A. R. 1968. Measurement of specific anosmia. Perceptual Motor Skills, 26, 143–164.

Amoore, J. 1979. Directions for preparing aqueous solutions of primary odorants to diagnose eight types of specific anosmias. Chemical Senses and Flavour, 4, 153–161.

Angulo, O., Lee, H.-S. and O'Mahony, M. 2007. Sensory difference tests, over dispersion and warm-up. Food Quality and Preference, 18, 190–195.

Barbary, O., Nonaka, R., Delwiche, J., Chan, J. and O'Mahony, M. 1993. Focused difference testing for the assessment of differences between orange juice made from orange concentrate. Journal of Sensory Studies, 8, 43–67.

Basker, D. 1980. Polygonal and polyhedral taste testing. Journal of Food Quality, 3, 1–10.

Bi, J. and Ennis, D. M. 2001a. Statistical methods for the A-Not A method. Journal of Sensory Studies, 16, 215–237.

Bi, J. and Ennis, D. M. 2001b. The power of the A-Not A method. Journal of Sensory Studies, 16, 343–359.

Bi, J. 2006. Sensory Discrimination Tests and Measurements: Statistical Principles, Procedures and Tables. Blackwell Publishing Professional, Ames, IA.

Brockhoff, P. B. 2003. The statistical power in difference tests. Food Quality and Preference, 14, 405–417.

Brockhoff, P. B. and Christensen, R. H.B. 2009. Thurstonian models for sensory discrimination tests as generalized linear models. Journal of Food Quality and Preference, doi:10.1016/j.foodqual.2009.04.003.

Dacremont, C., Sauvageot, F. and Ha Duyen, T. 2000. Effect of assessors expertise level on efficiency of warm-up for triangle tests. Journal of Sensory Studies, 15, 151–162.

Ennis, D. M. and Mullen, K. 1986. Theoretical aspects of sensory discrimination. Chemical Senses, 11, 513–522.

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

Frijters, J. E. R. 1979. Variations of the triangular method and the relationship of its unidimensional probabilistic model to three-alternative forced choice signal detection theories. British Journal of Mathematical and Statistical Psychology, 32, 229–241.

Frijters, J. E. R. 1984. Sensory difference testing and the measurement of sensory discriminability. In: J. R. Piggott (ed.), Sensory Analysis of Food. Elsevier Applied Science Publications, London, pp.117–140.

Frijters, J. E. R., Kooistra, A. and Vereijken, P. F.G. 1980. Tables of d' for the triangular method and the 3-AFC signal detection procedure. Perception and Psychophysics, 27, 176–178.

Gacula, M. C. and Singh, J. 1984. Statistical methods in food and consumer research. Academic, Orlando, FL.

Harris, H. and Kalmus, H. 1949. The measurement of taste sensitivity to phenylthiourea. Annals of Eugenics, 15, 24–31.

Huang, Y-T., and Lawless, H. T. 1998. Sensitivity of the ABX discrimination test. Journal of Sensory Studies, 13, 229–239; 8, 229–239.

Kim, H.-J., Jeon, S. Y., Kim, K.-O. and 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.

Kraemer, H. C. and Thiemann, S. 1987. How many subjects: Statistical power analysis in research. Sage, Newbury Park, CA.

Lawless, H. T. 1980. A comparison of different methods used to assess sensitivity to the taste of phenylthiocarbamide (PTC).Chemical Senses, 5, 247–256.

Lee, H-S., and O'Mahony, M. 2007. The evolution of a model: A review of Thurstonian and conditional stimulus effects on difference testing. Food Quality and Preference, 18, 369–383.

Liggett, R. A. and Delwiche, J. F. 2005. The beta-binomial model: Variability in over- dispersion across methods and over time. Journal of Sensory Studies, 20, 48–61.

Macmillan, N. A., Kaplan, H. L. and Creelman, C. D. 1977. The psychophysics of categorical perception. Psychological Review, 452–471.

Macmillan, N. A. and Creelman, C. D. 1991. Detection Theory: A User's Guide. University Press, Cambridge, UK.

Macrae, A. W. 1995. Confidence intervals for the triangle test can give reassurance that products are similar. Food Quality and Preference, 6, 61–67.

Mata-Garcia, M., Angulo, O. and O'Mahony, M. 2007. On warm-up. Journal of Sensory Studies, 22, 187–193.

Meilgaard, M., Civille, C. V., and Carr, B. T. 2006. Sensory Evaluation Techniques, Fourth Edition. CRC, Boca Raton, FL.

Mitchell, J. W. 1956. The effect of assignment of testing materials to the paired and odd position in the duo-trio taste difference test. Journal of Food Technology, 10, 169–171.

Nicolle, C. 1932. Biologie de l"Invention Alcan Paris, quoted in Beveridge, W. I.B. 1957. The Art of Scientific Investigation, Third Edition. Vintage Books, New York. p. 37.

O'Mahony, M. and Goldstein, L. R. 1986. Effectives of sensory difference tests: Sequential sensitivity analysis for liquid food stimuli. Journal of Food Science, 51, 1550–1553.

O'Mahony, M., Wong, S. Y. and Odbert, N. 1986. Sensory difference tests: Some rethinking concerning the general rule that more sensitive tests use fewer stimuli. Lebensmittel Wissenschaft und Technologie, 19, 93–95.

O'Mahony, M., Thieme, U. and Goldstein, L. R. 1988. The warm-up effect as a means of increasing the discriminability of sensory difference tests. Journal of Food Science, 53, 1848–1850.

Pangborn, R. M. and Dunkley, W. L. 1966. Sensory discrimination of milk salts, nondialyzable constituents and algin gum in milk. Journal of Dairy Science, 49, 1–6.

Peryam, D. R. 1958. Sensory difference tests. Journal of Food Technology, 12, 231–236.

Peryam, D. R. and Swartz, V. W. 1950. Measurement of sensory differences. Food Technology, 4, 390–395.

Pierce, J. R. and Gilbert, E. N. 1958. On AX and ABX limens. Journal of the Acoustical Society of America, 30, 593–595.

Roessler, E. B., Pangborn, R. M., Sidel. J. L. and Stone, H. 1978. Expanded statistical tables for estimating significance in paired-preference, paired difference, duo-trio and triangle tests. Journal of Food Science, 43, 940–941.

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

Schlich, P. 1993. Risk tables for discrimination tests. Journal of Food Quality and Preference, 4, 141–151.

Smith, G. L. 1981. Statistical properties of simple sensory difference tests: Confidence limits and significance tests. Journal of the Science of Food and Agriculture, 32, 513–520.

Stone, H. and Sidel, J. L. 1978. Computing exact probabilities in sensory discrimination tests. Journal of Food Science, 43, 1028–1029.

Stone, H. and Sidel, J. L. 2004. Sensory Evaluation Practices, Third Edition. Academic, Elsevier, New York.

Thieme, U. and O'Mahony, M. 1990. Modifications to sensory difference test protocols: The warmed-up paired comparison, the single standard duo-trio and the A, not-A test modified for response bias. Journal of Sensory Studies, 5, 159–176.

Whiting, R., Murray, S., Ciantic, Z. and Ellison, K. 2004. The use of sensory difference tests to investigate perceptible colour-difference in a cosmetic product. Color Research and Application, 29, 299–304.