EXAMINATION OF THE SUBJECT-DEFINED 2-AFC

A Thesis

Presented to the Faculty of the Graduate School

Of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Scott McClure

January 2008

ABSTRACT

Both the triangle test and 2-AFC are commonly used discrimination tests used in research and industry. The 2-AFC is statistically more powerful than the triangle test, but can only be used when the quality and direction of difference is known (i.e. one sample is sweeter). By allowing subjects to define their own criteria for use in the 2-AFC, it has been suggested, but not proven, that the 2-AFC procedure can be used when the sensory difference between two samples is not known or easily defined. This modified 2-AFC procedure is called the Subject Defined 2-AFC (SD-2-AFC). This study compares the SD-2-AFC to the triangle and conventional 2-AFC (2-AFC) under a number of realistic conditions. Four food systems, different in the magnitude and quality of sensory difference, were examined. Results demonstrated that, in conditions where the 2-AFC could have been used, the SD-2-AFC did not perform as well as either the 2-AFC or triangle tests. In conditions where the 2-AFC could not have been used, the SD-2-AFC did not perform as well as the triangle test. The failure of the SD-2-AFC was due to subjects inverting their criteria (i.e., picking 'less sweet' instead of 'sweet'). Inverted subjects performed lower than expected, whereas non-inverted subjects performed better than the 2-AFC or triangle tests. These results were explained in terms of signal detection theory.

BIOGRAPHICAL SKETCH

Scott McClure completed his high school education in Fairfax, Virginia. During this time he was accepted into the Virginia Governor's School for Agriculture, a summer program conducted at Virginia Tech's campus where he was formally introduced to the field of food science. After graduating one of his high school's Valedictorians in the class of 2004, Scott enrolled at Cornell University in the College of Agriculture and Life Science to major in food science as part of the class of 2008. Upon enrollment, Scott was accepted into the Hunter T. Rawlings Cornell Presidential Research Scholar Program (then named the Cornell Presidential Research Scholar Program). This program, in addition to providing a loan reduction, provided him with a significant research grant to encourage undergraduate research. As part of the program's requirements, he interviewed several professors to decide where to do research. As a result of these interviews, he met Dr. Harry T. Lawless in the Food Science Department's Sensory lab. For the rest of his undergraduate career, Scott conducted research projects with Dr. Lawless focusing on different aspects of metallic taste. Through this research, Scott was able to coauthor two papers and present at the ACHEMS annual conference twice. Scott was able to graduate with his Bachelor's Degree a semester early in the winter of 2007. He then continued on as a graduate student in Dr. Lawless' lab, during which time this thesis was produced.

ACKNOWLEDGEMENTS

I would like to first thank both the Hunter T. Rawlings Cornell Presidential Research Program and the Cornell University Department of Food Science for their assistance in the funding of this research and my Master's program. Thank you to Dr. Martin Wiedmann, the Director of Graduate Studies in the Food Science Department, for his assistance in finding a way to fund my Master's program. Thank you to Janette Robbins for her support and guidance throughout my program. Thank you to Kathy Chapman for her assistance in the Sensory Lab. Thank you to Mike Nestrud and Effie Epke for the motivation they provided during my research. Thank you to Dr. Bruce Halpern for his assistance as my minor advisor. A final thank you goes to Dr. Harry T. Lawless. Not only for his guidance as my major advisor, but for the four years of mentorship that he provided during both my undergraduate and graduate education. This mentorship, more than providing me with publications and degrees, changed the way I approach scientific problems and will impact the way I think throughout my career. I would not have achieved what I have without his guidance. Thank you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

LITERATURE REVIEW

1.1 INTRODUCTION TO SENSORY EVALUATION OF FOODS

Sensory evaluation draws from the fields of psychology, statistics, and biology with the purpose of using human subjects to determine 1) if two or more samples are the same or different, 2) to what degree two or more samples differ in one or more sensory dimensions, and 3) if two or more samples differ in desirability. Sensory evaluation has broad applications in a number of industries as well as in academia. In industry, sensory evaluation can be applied to textiles, fragrances, foods, and any number of other consumer goods. Sensory evaluation of foods can guide product developers wishing to make a new product or improve a current product, provide a cost reduction without changing a current product, identify a lapse in quality control, direct developers in fixing a quality control issue, and more. In academia, sensory evaluation of foods can be applied to pure research of the effects of compounds on the senses or the senses themselves. It is also used to determine the sensory effects of innovative processes/ingredients/equipment aimed to increase the safety, nutrition, shelf life, and functionality of foods. Often, sensory evaluation determines if a research idea is developed enough to cross into industrial use.

The "central dogma" of sensory evaluation is that the test must match the objective (Lawless and Heymann, 1998a). Due to this, a multitude of methods are used depending on the type of question being asked as well as the type of answer desired. Tests used in sensory evaluation of food can be divided into three main types within two main categories: analytic tests- consisting of discrimination and descriptive tests, and hedonic or affective tests (Lawless and Heymann, 1998a). Discrimination tests are used to determine if two or more products are different. Descriptive tests are

used to determine how two or more products are different. Affective tests are used to determine which of two or more products is most preferred.

The subjects used for one type of test are not always suitable for another type of test. Discrimination test subjects must have sensitivity equal to the sensitivity of the answer sought. If one is researching the limits of the senses, they should be extremely sensitive. If one is researching the population as a whole, they should be of average sensitivity. Descriptive test subjects are often trained to the point where they can be considered similar to instruments. Affective test subjects must have the same sensitivity as the intended consumers of the product. These differences are important because it is usually possible to find a panel naïve enough to miss a very large difference as well as a panel experienced enough to find a minute one. It is only when the purpose of the test is known that a suitable panel can be chosen.

Though based in solid models and theory, sensory evaluation is nothing if not practical. The methods and analyses are constantly adapted to the realities of testing in a world of finite time, funding, and subjects. Research is constantly conducted to find procedures and analyses that can find the same answers faster, cheaper, and requiring fewer subjects than those currently used. The exact combination of methods and analysis used by a company is protected like any other trade secret thought to give them an edge on the competition. As they are generally the first tests used and provide the most general answers, the purpose of this review is to explain where the research stands on procedures and analyses for discrimination tests used in the sensory evaluation of foods. Multiple texts exist for those seeking an introduction into sensory evaluation beyond the scope of this review (Lawless and Heymann, 1998a; Piggott, 1988).

1.2 DISCRIMINATION TESTS

Discriminations tests are often the first, and sometimes only, test used when examining a problem (Frijters, 1988; Lawless and Heymann, 1998c). It is logical that the nature of differences between products cannot be determined if no noticeable difference exists between them. These tests are also inexpensive compared to descriptive tests, which often require extensive training periods, and affective tests, which often require extensive recruiting of users and/or likers of that product. (How many people do you know that eat at two or more frosted strawberry-filled breakfast pastries a week?) Table 1 presents a summary of the major discrimination tests grouped by the theoretical cognitive strategy they promote (discussed in 1.3.1). This section describes the procedures of these tests in detail. Except where otherwise cited; procedures in this section can be found in the text by Lawless and Heymann (1998). All of these tests compare varying numbers of two different stimuli, A and B.

Peryam and Swartz (1950) first described the duo trio procedure. In the duo trio, subjects are given one reference sample followed by two test samples A and B. Subjects are told to match the reference sample to the correct test sample. Within a duo trio, all subjects may receive the same reference. This is known as the constant reference duo trio. If half the subjects receive A as the reference and half receive B as the reference it is known as the balanced reference duo trio. Peryam and Swartz (1950) also included an initial presentation of the reference sample to warm up subjects. As discussed later, this single sample presentation would be best described as a preview sample rather than a warm up. The ABX is the inverse of the duo trio (Huang and Lawless, 1998). Subjects first receive two reference samples and then either test samples A or B. Subjects then match the test sample to one of the reference samples. In the dual standard, subjects receive both reference samples and both test

Table 1. Summary of common discrimination test procedures. A and B refer to different test samples, RA and RB refer to reference samples of the same products. *More combinations exist than were mentioned (e.g. the triangle can be A, A', B or B, B', A and in any random order).

Cognitive Strategy	Test Name(s)	Samples Presented*	Task	Chance Level
Comparison of Distances	Duo Trio	RA, A, B	Match A to RA	1/2
	ABX	RA, RB, A	Match A to RA	1/2
	Dual Standard	RA, RB, A, B	Match A to RA and B to RB	1/2
	Triangle, Three Interval Oddity	A, B, B'	Group into A and B, B' (Pick odd Sample)	1/3
	Tetrad - Difference	A, A', B, B'	Group into A, A' and B, B'	1/6
	Two out of Five Sorting	A, A', B, B', B"	Group into A, A' and B, B', B"	1/10
	Four out of Eight Sorting, Harris-Kalmus	A, A', A", A"', B, B', B", B"'	Group into A, A', A", A"' and B, B', B", B"'	1/70
	Four Interval AX, Dual Pair, 4IAX	A, B and A, A'	Pick Pair A, B	1/2
	X out of Y Sorting	A, … A_X, B, …B_{Y-X}	Group into A's and B's	$\frac{Y!}{X!(Y-X)!}$
Skimming	Tetrad - Directional	A, A', B, B'	Group into A, A' and B, B' (Told how B is different)	1/6
	Two Alternative Forced Choice, 2-AFC, Paired Comparison, Pair	A, B	Pick B (Told how B is different)	1/2
	Three Alternative Forced Choice, 3-AFC, Directional Triangle	A, A', B	Pick B (Told how B is different)	1/3
	n-Alternative Forced Choice, n-AFC, m-AFC	A,…A_{n-1}, B	Pick B (Told how B is different)	1/n
Tau (Beta)	Same/Different, Simple Difference	A, B or A, A' or B, B'	Identify as Same or Different	1/2
	A, Not A (Traditional)	RA then A or B	Identify as A or not A	1/2
	A, Not A (R-Index)	RA then A or B	Identify as 'Definitely A', 'Maybe A', 'Maybe Not A', or 'Definitely not A'	N/A

samples (O'Mahony et al., 1986). Subjects then match the test samples to the reference samples.

In the triangle, subjects receive two identical samples and one odd sample (Helm and Trolle, 1946). They are instructed to identify the odd sample. In the tetrad, subjects receive four samples, two from one stimulus and two from the other (Delwiche and O'Mahony, 1996). Subjects then must sort them into matching pairs (Delwiche and O'Mahony, 1996). In the directional tetrad, subjects are told how the samples are different (i.e. sweetness) (Delwiche and O'Mahony, 1996). In the difference tetrad, subjects are only told that the samples are different (Delwiche and O'Mahony, 1996). In the 2/5, subjects are given five samples with two from one stimulus and three from the other. Subjects must then sort them correctly into one group of two and one group of three (Amoore et al., 1968; Amoore, 1977). In the 4/8, subjects are given eight samples with four from one stimulus and four from the other (Harris and Kalmus, 1949). Subjects must then sort them correctly into two groups of four.

In the 4IAX, subjects receive two pairs of samples (Rousseau and Ennis, 2001). One pair consists of identical stimuli; the other consists of different stimuli (Rousseau and Ennis, 2001). Subjects must indicate which pair contains the different stimuli (Rousseau and Ennis, 2001). In the same/different test, subjects receive one pair of stimuli, either the same or different from each other. Subjects must indicate if the two samples in the pair are the same or different from each other. This may be a completed block design, where subjects see both pairs, or an incomplete block design, where subjects see only one pair. In the A-not-A test, subjects receive a reference sample and then either sample A or sample B. Subjects then determine if the test sample is the A sample or not the A sample. Inclusion of certainty ratings is conceptually similar to the degree of difference test proposed by Aust et al. (1985). In

the degree of difference test, subjects receive a reference sample followed by a test sample. They then rate the test sample on a scale with 'no difference' on one extreme and 'extremely large difference' on the other extreme (Aust et al., 1985). The degree of difference test provides a method to test heterogeneous products (such as stews) (Aust et al., 1985).

In the 2-AFC, subjects receive two different samples. Subjects must indicate the sample that matches the difference indicated (i.e. sweetest). In the 3-AFC, two identical samples and one odd sample are provided. Subjects must indicate the sample with the most or least of the difference indicated. Selecting the odd sample is considered a correct response. The n-AFC is the same, except that there are n-1 of the identical samples.

These tests make up the discrimination section of the "sensory toolbox" (Lawless and Heymann, 1998c). Just as a hammer is neither better nor worse than a screwdriver, each of these tests have certain characteristics that make them better in some situations in worse in others. The following sections will discuss the factors and theories that explain the differences among these procedures.

1.3 SIGNAL DETECTION THEORY

Good sensory practices, such as randomization of presentation orders and rinsing between samples are designed to minimize effects such as fatigue, desensitization, sensitization, response bias, attention, etc. and reduce variance in order to make sensory tests more repeatable (Lawless and Heymann, 1998b). However, even in a hypothetical "perfect" sensory test where all of these effects have been eliminated, a certain amount of variance is still expected (Thurstone, 1927). Signal detection theory (SDT) provides a method to account for this variance. The text by Green and Swets (1966) provides a thorough explanation of signal detection theory.

Signal detection theory has been used to explain how humans are able to distinguish a taste, noise, smell, etc (signal) from its surroundings (noise). Based on Thurstone's "law of comparative judgment" (1927), SDT assumes that sensory experiences fall on a continuum, rather than being binomial in nature (Frijters, 1979; Lawless and Heymann, 1998d). Given a constant stimulus, repeated exposures will generate a range of sensations normally distributed around a mean sensation (Frijters, 1979). This distribution is known as the 'signal distribution' (Lawless and Heymann, 1998d). A second normal distribution, with equal variance (in some models) to the first, overlaps the signal and is known as 'noise distribution' (Lawless and Heymann, 1998d). Several authors have provided figures visualizing the concepts of SDT similar to those used in this paper (Lawless and Heymann, 1998d; O'Mahony and Rousseau, 2002; O'Mahony, 1995; O'Mahony et al., 1994; Rousseau, 2001; Frijters, 1979). For threshold testing, this 'noise' distribution can be thought of as the sensations from a "blank stimulus, such as distilled water (Lawless and Heymann, 1998d). Figure 1 presents a SDT interpretation of a comparison of the intensity two samples on a single sensory dimension (X-Axis) or the comparison of signal to noise. The curves represent the sensations generated by the noise / weaker sample (left hand curve) and the signal / stronger sample (right hand curve) The distance between the peaks represents sensory difference between the samples (δ or its estimate d'). The two graphs represent (A) a relatively small sensory difference and (B) a relatively large sensory difference. The goal is to estimate the true sensory difference (using d') from the test data.

SDT can also be applied to discrimination tests (Lawless and Heymann, 1998d). Instead of visualizing the two overlapping curves as the perceptual signal and noise from one sample and its background, they are visualized as the sensations that

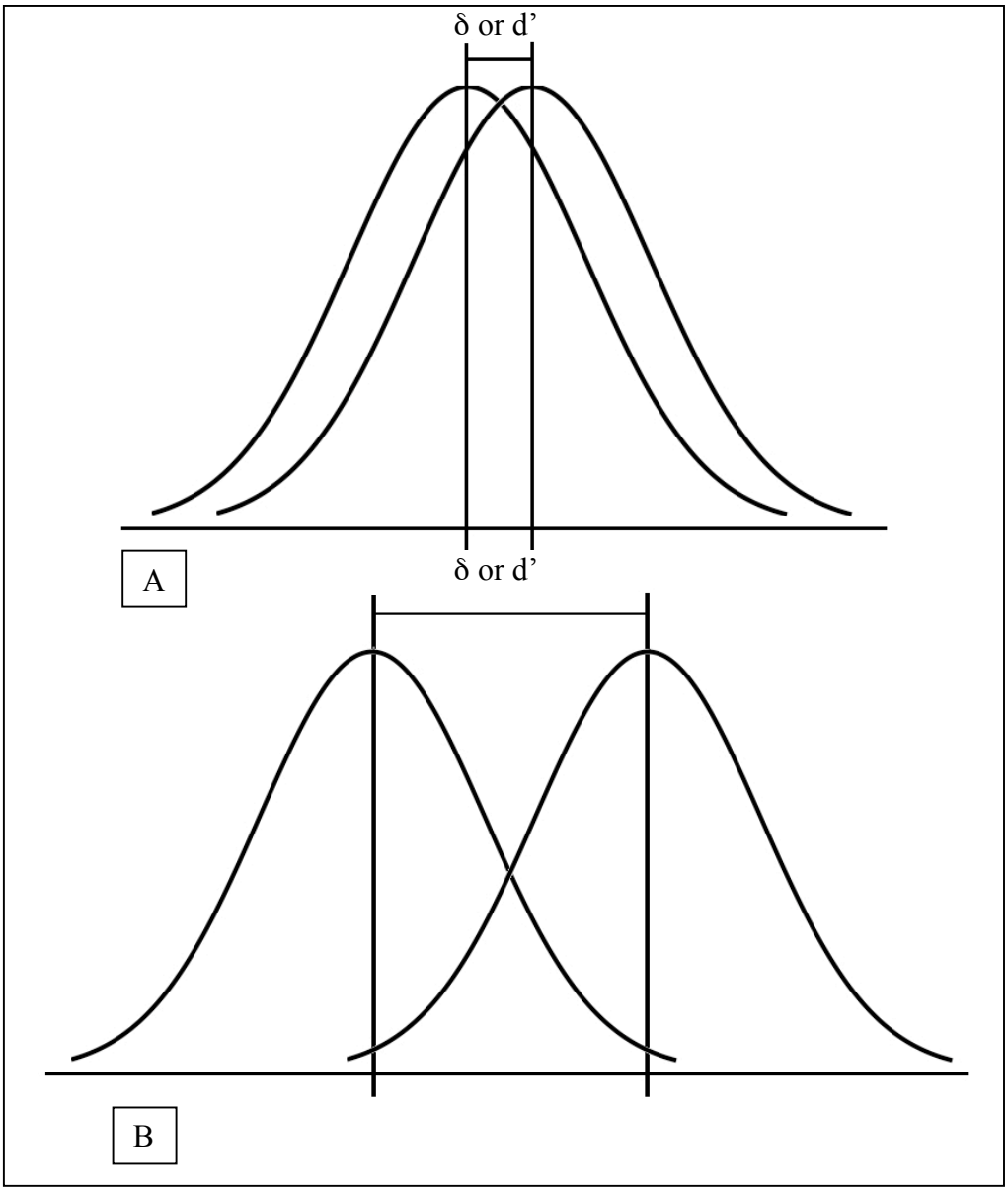


Figure 1. Signal Detection Theory interpretation of a comparison of the intensity two samples on a single sensory dimension (X-Axis). The distance between the peaks represents sensory difference between the samples (δ or its estimate, d'). The two graphs represent (A) a relatively small sensory difference and (B) a relatively large sensory difference.

arise from two different samples (Lawless and Heymann, 1998d). The closer the two peaks of these distributions, the more confusable the stimuli (Frijters, 1979).

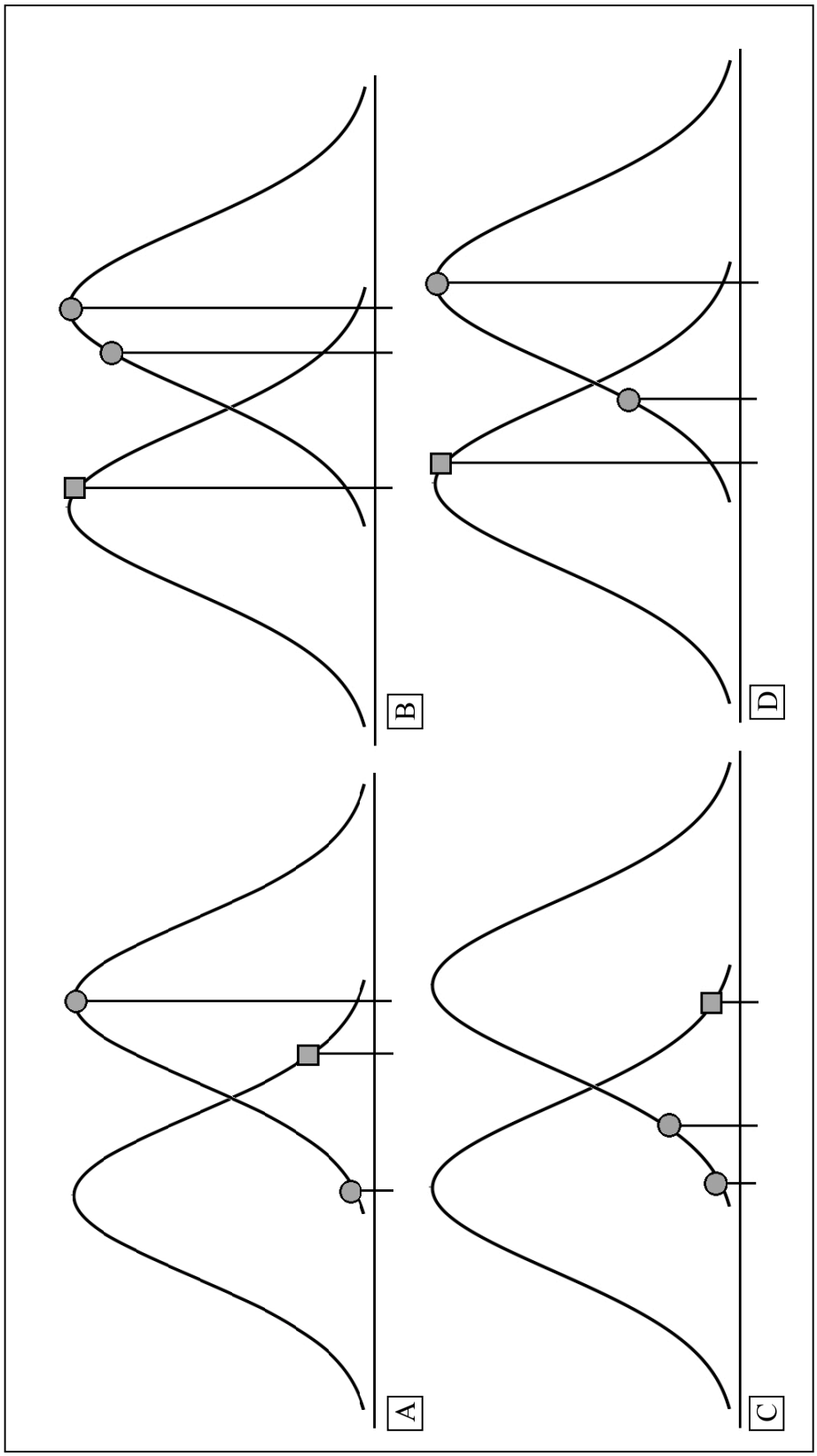
1.3.1 RESPONSE BIAS AND COGNITIVE STRATEGIES

The unique psychological condition of a particular subject at the moment of response has an effect on the answer they give. For example, a subject who believes that a researcher would not give them two identical samples very often is predisposed to calling two samples different in the same/different procedure. This effect, known as response bias or criterion variation, has been described as "the central problem in discrimination testing" (O'Mahony and Rousseau, 2002).

Many sources of bias can be controlled for by good sensory practices (Lawless and Heymann, 1998b). These prevent a subject from being predisposed to picking a sample due to it having a meaningful label (such as a letter or a single number) (Lawless and Heymann, 1998b). Using different numbers for all samples that a single subject will see prevents subjects from remembering or writing down numbers between sample sets (Lawless and Heymann, 1998b). Standard sample sizes, cups, and presentation styles reduce bias as well (Lawless and Heymann, 1998b). When all good sensory practices are used, response bias can still exist (O'Mahony and Rousseau, 2002).

Response bias is dependent on the specific cognitive strategy used by subjects in specific testing procedures. When subjects taste samples, experiences can be visualized as specific points along a single dimension (Figure 2). These points are drawn as samples from the normal distributions implied by SDT (Figure 2). Subjects then choose a response based on the positions of the samples on the dimension. The different methods subjects use to pick the correct sample are known as cognitive strategies or decision rules (Frijters, 1979). Four main cognitive strategies exist: τ-criterion, β-criterion, skimming, and comparison of distances (Ennis et al., 1988;

Figure 2. Signal Detection Theory interpretation of a comparison of the intensity two samples on a single sensory dimension (X-Axis) in a three-sample test (Triangle or 3-AFC) where two stronger and one weaker sample are provided. The circles and squares each represent a single sensory experience from a stronger and weaker distribution, respectively. With the comparison of distances strategy, subjects identify the sample that is furthest from the other two samples. With the skimming strategy, subjects identify the sample that is the highest (furthest to the right) in the specified sensory dimension (X-Axis). The four figures represent the four possible cases where both skimming and comparison of distances would be incorrect (A), where both would be correct (B), where only comparison of distances would be correct (C) and where only skimming would be correct (D).


A
B
C
D

Frijters, 1979; Rousseau, 2001; O'Mahony et al., 1994). Guessing is a cognitive strategy as well. If the other cognitive strategies are correct, subjects should only guess when the momentary sensations elicited by the samples appear identical (i.e., Figure 3). Each discrimination testing procedure uses (is designed so that subjects will likely use) of one or more of these strategies (Rousseau, 2001).

The τ-criterion applies to same/different judgments. It is the minimum magnitude of difference required for subjects to call two samples different or the maximum magnitude of difference required for subjects to call two samples the same. It can be visualized as a line segment parallel to a sensory continuum (Figure 4) (O'Mahony and Rousseau, 2002). Subjects experience both samples along this continuum (O'Mahony and Rousseau, 2002). If the samples are closer together than the length of the line segment, they are described as 'the same' (O'Mahony and Rousseau, 2002). If the samples are further apart than the length of the line segment, they are described as 'different' (O'Mahony and Rousseau, 2002). The length of this line segment is expected to be different for each subject, and either fixed or constant for each rating by a single subject (O'Mahony and Rousseau, 2002; Ennis et al., 1988). For example, if a subject performs four tests in a row and selects 'different' three times in a row, they might expect at least one of the set to be 'same' and expand the length of their τ-criterion. Use of a τ-criterion requires that subjects are comparing the similarity of two samples as opposed to comparing one sample to a reference (Rousseau, 2001).

The comparison of distances cognitive strategy is used when three or more samples are being compared on an unknown sensory dimension (Frijters, 1979). Subjects taste the samples and decide what dimension to compare them on (O'Mahony et al., 1994). This dimension could represent a simple taste quality (such as sweetness), a combination of taste qualities, overall differences, or something abstract.

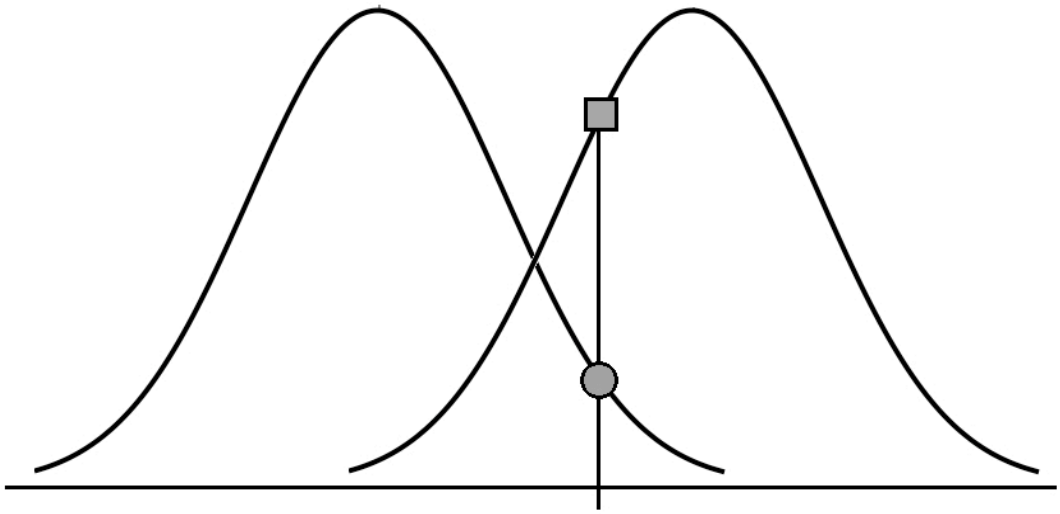


Figure 3. Signal Detection Theory interpretation of a comparison of the intensity two samples on a single sensory dimension (X-Axis) in a two-sample test (2-AFC) where one stronger and one weaker sample are provided. The circles and squares each represent a single sensory experience from a stronger and weaker distribution, respectively. In this example the subject would be expected to guess at random.

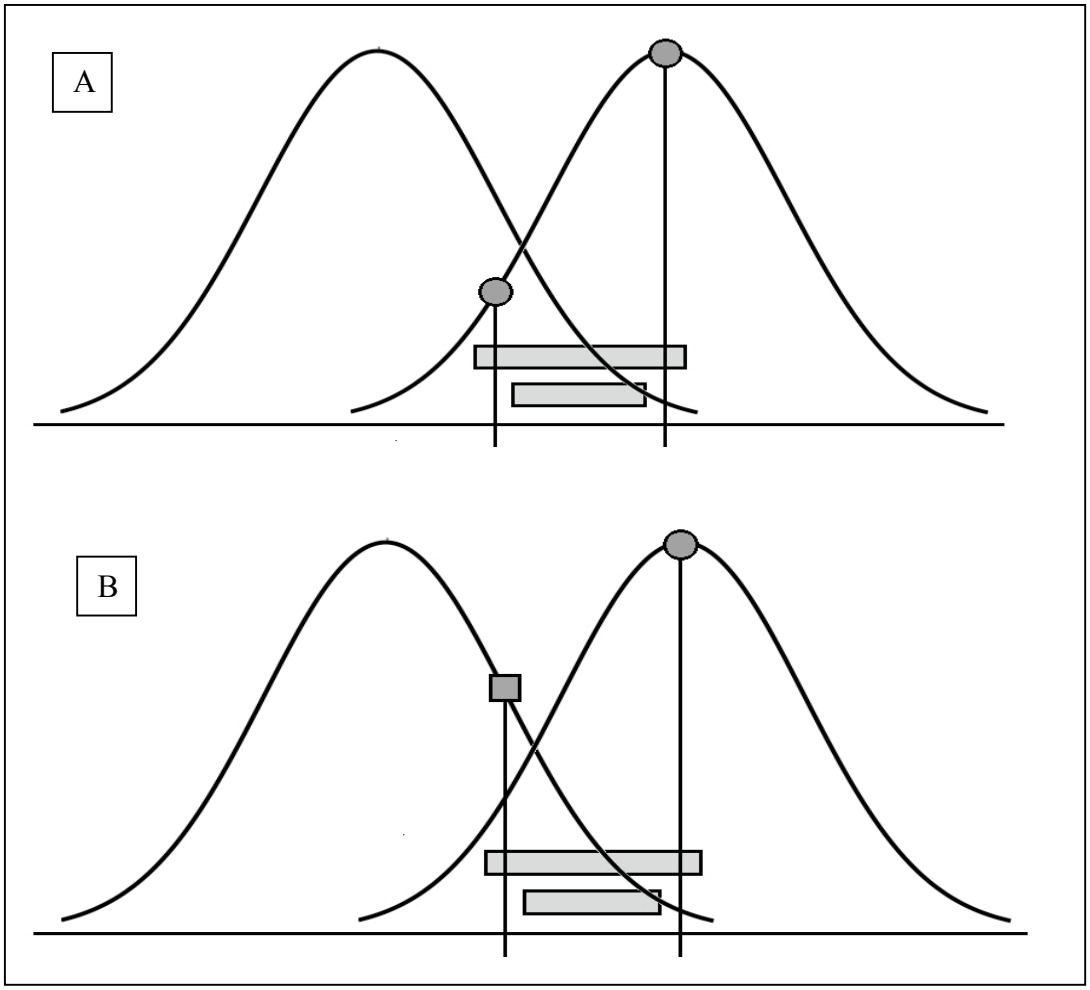


Figure 4. Signal Detection Theory representation of the τ-criterion. The circles and squares each represent a single sensory experience from a stronger and weaker distribution, respectively. Grey rectangles represent a large and small τ-criterion. The two sets of graphs (A and B) represent situations where the same criterion could result in correct or incorrect answers.

The distances between the samples are then compared. Because these distances are determined by the properties of the samples, not the psychological state of the subjects, the comparison of distances strategy is not considered vulnerable to response bias (Bi, 2006b).

Tests that use the comparison of distances strategy have differences between each other. These differences depend on the number of comparisons that must be made. In the case of the ABX, only two comparisons must be made. Since the two reference samples are described as being different from each other in the instructions of the test, the only comparisons that need to be made are RA to A and RB to A. The pair with shortest distance is chosen as the same. Similarly in the duo trio, since the two test samples are assumed to be different, they do not need to be compared. In the case of the triangle, three comparisons must be made: A to A', A to B, and A' to B. The pair with the shortest distance is chosen as the same or the sample with the two longest distances is chosen as different.

The case of the dual standard is less clear-cut. It is possible that subjects treat it as an ABX. In this scenario, subjects would match one test sample to the reference and simply assume the other matches the second reference. The alternative would be to make four comparisons: RA to A, RA to B, RB to A, and RB to B. The pair with the shortest distance would be matched and the other two matched by default. The difference tetrad and the sorting methods (2/5, 4/8, X/Y) all have a large number of comparisons (the additive factorial of one minus the total number of samples).

The skimming strategy can be used when the dimension and magnitude of the difference is specified (Frijters, 1979). As with the comparison of distances strategy, samples are tasted and placed along a sensory dimension. Subjects then identify the sample highest or lowest (depending on the instructions) along the dimension. In the case of the directional tetrad, the two samples highest or lowest (depending on the

instructions) in the dimension are chosen. In the 2-AFC, the one sample highest or lowest (depending on the instructions) in the dimension is chosen.

The β-criterion can be visualized as a line somewhere along a sensory continuum (Figure 5). This sensory continuum is defined by the subject in response to either directions or experience with the samples (i.e., reference sample in the A, Not A) (Rousseau, 2001). In the A, not A procedure subjects select a dimension that explains the difference between the A and not A samples, sweetness for example (O'Mahony and Rousseau, 2002; Rousseau, 2001). A β-criterion is then placed along this dimension (Figure 5) (Rousseau, 2001). In Figure 5, sensations on the left of the β-criterion would be called 'not sweet' and sensations on the right of the line would be called 'sweet'. If the A sample was defined as 'sweet' by the subject, then sensations on the left of the β-criterion would be called 'not A' and sensations on the right of the line would be called 'A' (Rousseau, 2001). The placement of the β-criterion is expected to vary similarly to the τ-criterion (O'Mahony, 1995). For example, if a subject performs four tests in a row and selects 'sweet' three times in a row, they might expect at least one of the set to be 'not sweet' and shift their β-criterion to the left. Use of a β-criterion generally requires that subjects are comparing a single sample to a sensory quality (Rousseau, 2001). That is, a simple yes/no response to a single stimulus presentation (Figure 5). However, it has been suggested that subjects could use a β-criterion for multiple sample tests (Rousseau, 2001).

It has been suggested that, under conditions where subjects become familiar with the stimuli used, a β-criterion may be used instead of a τ-criterion (Rousseau, 2001). In these conditions, subjects assume that there are only two types of stimuli and learn how they are different (Rousseau, 2001). These differences are used to draw a β-criterion (Rousseau, 2001). Similarly, it could be possible for subjects to predict the important sensory dimension or a limited number of important dimensions after

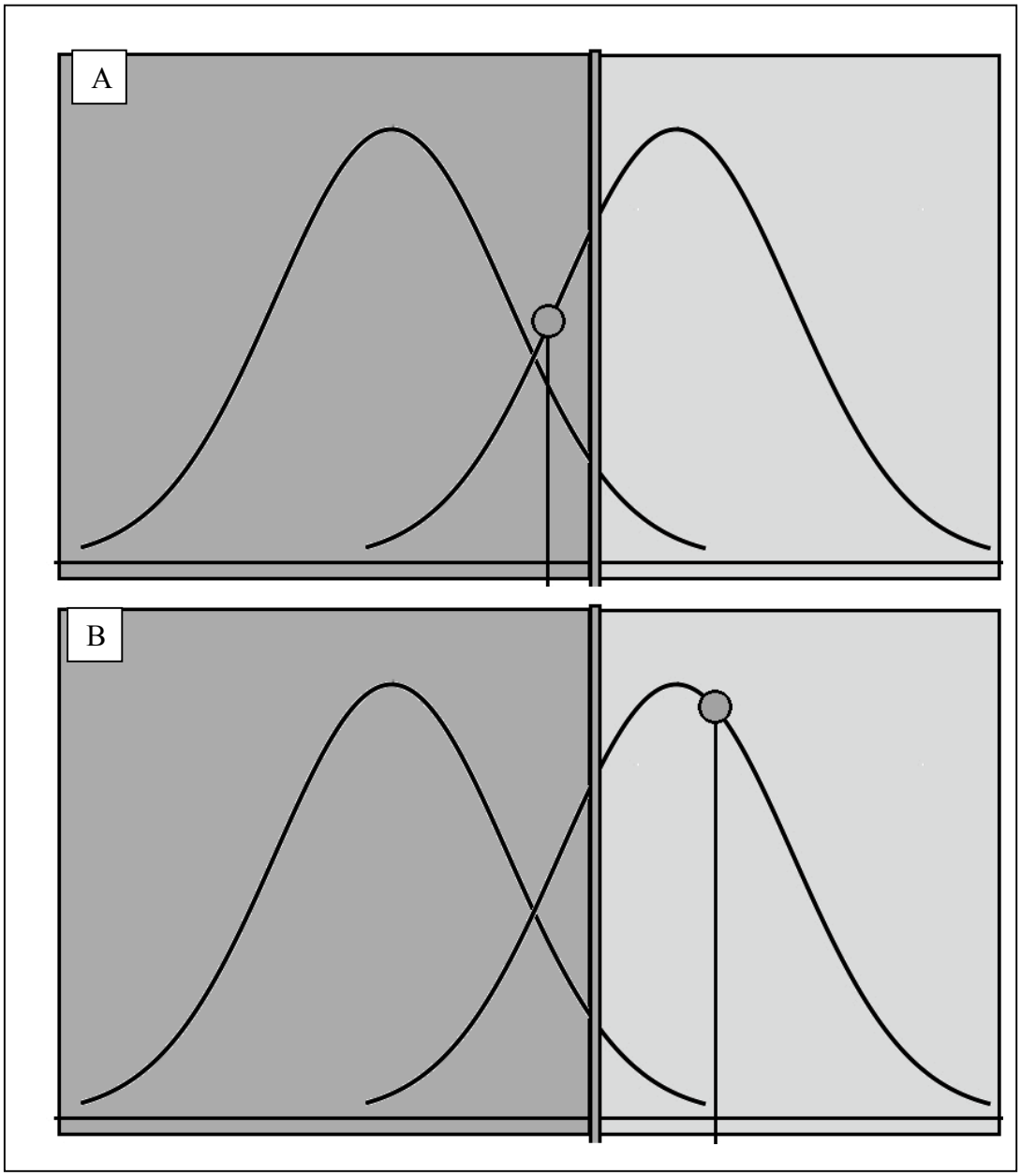


Figure 5. Signal Detection Theory representation of the β-criterion. The circles each represent a single sensory experience from the stronger distribution. Grey rectangles represent the same β-criterion in situations where it would result in (A) an incorrect and (B) a correct discrimination.

tasting the first sample (especially if the stimuli are simple). These dimensions could then be used to draw one or many β-criteria. This would be most likely in tests were samples are presented as references, since subjects know to focus on those samples.

It has been suggested that the β-criteria, or β-strategy, could be used instead of the comparison of distances strategy (Figure 6) (Rousseau, 2001). In this scenario subjects would choose one or many sensory dimensions to rate the samples on, each with their own β-criterion (Rousseau, 2001). In instances where the β-criterion divides the three samples into a group of two and a group of one, the odd sample would be chosen (Rousseau, 2001). In instances where all three samples are on one side of the β-criterion several things could happen. Subjects could then switch to an alternative β-criterion (Rousseau, 2001). This is more likely to be the case in complex stimuli where more than one attribute has been changed (Rousseau, 2001). Knowing which cognitive strategy subjects are truly using is critical for correctly analyzing and interpreting the raw data from a discrimination test, as the next section discusses.

1.4 STATISTICAL ANALYSIS OF DISCRIMINATION TESTS

1.4.1 POWER AND SENSITIVITY

Regardless of the discrimination test, the form of the data will be the same. Each test that each individual subject performs is viewed as a binomial success or failure, a '1' or a '0' (Bi, 2006b). The sensory scientist must then decide the best way to analyze and interpret these data points. This section will look at these analyses in terms of their ability to compare results from different procedures.

The goal of statistical analyses on discrimination tests is to interpret the data in a way that identifies true differences and ignores spurious ones. A "true difference" can be defined as a difference that the population of interest would be able to detect. This means that for the same change in a food product, a true difference may or may not exist dependent simply on the motive of the test. For example, a cost reduction in

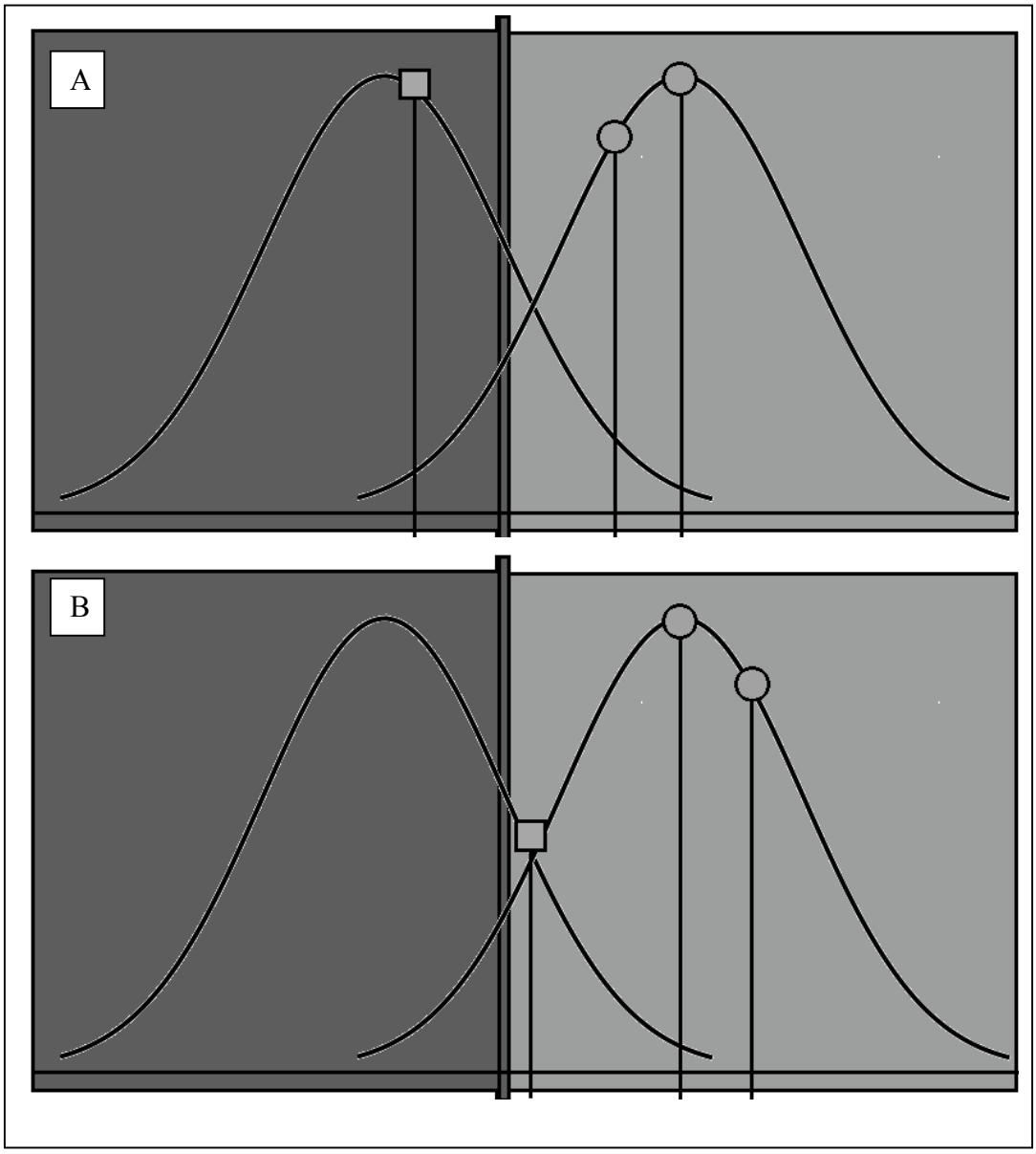


Figure 6. Signal Detection Theory representation of the β-strategy on a triangle test. The circles and squares each represent a single sensory experience from a stronger and weaker distribution, respectively. Grey rectangles represent the same β-criterion in situations where it would result in (A) a correct and (B) an unknown answer.

a coffee may result in a true difference for loyal consumers of that brand, but not for infrequent consumers. If the product is new and does not have a loyal following, the difference may not be important.

Power is the ability of the analysis to not miss a true difference (Lawless and Heymann, 1998f). Sensitivity can be thought of as the ability for subjects to correctly discriminate products in a test (Lawless and Heymann, 1998f). Power is mainly determined by the design of the statistical analysis, the number of observations, and the variance (Lawless and Heymann, 1998f). The training of the subjects themselves, the presence or lack of good sensory practices, and the testing procedure itself all may influence sensitivity (Lawless and Heymann, 1998f).

Equation 1 explains the relationship between α, β, and number of observations, N (Lawless and Heymann, 1998d). N is usually the number of subjects or judges, $p\text{-}p_a$ is a term representing the size of the sensory difference that needs to be detected, Z_α and Z_β are the Z-scores associated with the α and β-risks, respectively, of the test. The only way to increase statistical power for a test looking for a set difference on a specific set of samples, without increasing the risk of making an incorrect conclusion, is by increasing the number of subjects (Lawless and Heymann, 1998d) (Equation 1). Increasing the level of α-risk or the size of difference being investigated ($p\text{-}p_a$) can also increase power (Lawless and Heymann, 1998d). The larger the difference being examined, the more likely it will be found. However, if a researcher is interested in a specific difference, this may not be an option. Additionally, α-risk (the chance of finding a spurious difference) and β-risk (the chance of missing a true difference) are inversely proportional. Power is 1 – β-risk (Equation 1). This means that increasing the level of allowable α-risk will increase the power of the test. It is interesting to note that reducing α-risk either increases β-risk linearly or increases the number of subjects required exponentially. Increasing the number of observations can reduce both the α

and β level, but only in a square root fashion. Determining the power of a test is an important, though often overlooked, step in experimental design (Radkins, 1957; MacRae, 1995a).

$$N = \left\{ \frac{Z_{alpha}\sqrt{pq} + Z_{beta}\sqrt{p_a q_a}}{p - p_a} \right\}^2 \quad (1)$$

where N = Number of Observations (Subjects)

p = probability of correct decision (Null Hypothesis)

q = 1 - p

p_a = probability of correct decision (Alternative Hypothesis)

$q_a = 1 - p_a$

Z_{alpha} = Z-score associated with the level of α-risk

Z_{beta} = Z-score associated with the level of β-risk

$p - p_a$ = required level of discrimination

1.4.2 PROPORTION CORRECT BASED METHODS

The original, and still most frequent, method of analysis is to compare the results to what would be expected from subjects blindly guessing (Lawless and Heymann, 1998e). Blindly guessing means that each of the test samples has an equal likelihood of being chosen (Lawless and Heymann, 1998e). The data are transformed into the proportion correct and tested against the probability of getting a correct answer by chance (Table 1) (Lawless and Heymann, 1998e). This method takes the null hypothesis that the population proportion correct is equal to chance probability and the alternative hypothesis that the population proportion correct is greater than chance probability (Lawless and Heymann, 1998e). The results are assumed to fit a normal approximation to the binomial distribution (Lawless and Heymann, 1998e). Equation 2 can be used to determine the Z-score for a specific test. The Z-score is assumed to come from a one-tailed distribution, since values below chance are not

typically encountered. Tables can then be used to determine the likelihood of finding that Z-score by chance.

$$Z = \frac{(P_{obs} - p) - 1/2n}{\sqrt{pq/n}} \tag{2}$$

where P_{obs} = proportion of correct responses

n = total number of responses

p = probability of correct decision by chance

The null hypothesis is then rejected if the proportion correct is enough higher than the chance level ($p - p_a$) to allow no more than the accepted level of α-risk (Lawless and Heymann, 1998e). Subject sample size can be determined to ensure the desired power level (ability to call different samples different) as indicated in Equation 1 (Lawless and Heymann, 1998d).

Two methods have been used to compare the proportion correct from one method to another method with a different chance probability (Duo Trio vs. 3-AFC for example). The first method is to account for chance in the proportion correct using Abbot's formula (Equation 3) (Lawless and Heymann, 1998c). In this equation, $P_{observed}$ refers to the number of correct discriminations over the total number of tests. P_{chance} refers to the proportion correct expected if all subjects guessed randomly between the available options. This solution is known as the chance-adjusted proportion correct ($P_{adjusted}$).

$$P_{adjusted} = \frac{P_{observed} - P_{chance}}{1 - P_{chance}} \tag{3}$$

Another method is to determine the number of discriminators in a test (Lawless and Heymann, 1998c). This method assumes that there are two types of subjects: discriminators and non-discriminators (Equation 4.1) (Lawless and Heymann, 1998c). Non-discriminators have also been referred to as non-recognizers and discriminators as recognizers (Ferdinandus et al., 1970). Discriminators (D) are assumed to detect the difference and perform the test correctly (Lawless and Heymann, 1998c). Non-discriminators (XD) are assumed to not detect the difference and guess, performing the test correctly at the level suggested by chance (P_{chance}) (Equation 4.2) (Lawless and Heymann, 1998c). Both Equations 4.1 and 4.2 are derived from Equation 3 by replacing $P_{adjusted}$ with D/N.

$$N = D + XD \tag{4.1}$$

$$C = D + (P_{chance})XD \tag{4.2}$$

It is important to remember that this method is not meant to actually identify which subjects were discriminators and which were non-discriminators (Lawless and Heymann, 1998c). This method does provide an easy to understand output and is simple, fast, and easy to explain to non-sensory professionals. However, this method does not account for variations in cognitive strategy (Ennis, 1993; Rousseau and Ennis, 2007). Because of this, comparing results across tests with different cognitive strategies can result in faulty conclusions (Ennis, 1993; Rousseau and Ennis, 2007).

1.4.3 PARADOX OF THE NON-DISCRIMINATORY DISCRIMINATORS

Take the example of Byer and Abrams' (1953) comparison of the triangle and 3-AFC tests. Using the same subjects and comparing the same stimuli, it would be expected that the tests should produce the same results. This is not what occurs (Byer and Abrams, 1953). Under these conditions, subjects consistently perform better on the 3-AFC than on the triangle (e.g. Byer and Abrams, 1953; Hopkins and Gridgeman, 1955; Delwiche and O'Mahony, 1996; Rousseau and O'Mahony, 1997; Stillman,

1993; Tedja et al., 1994; Masuoka et al., 1995; Filipello, 1956). When this effect was first found, it was termed the 'paradox of the nondiscriminatory discriminators' because subjects that could discriminate two samples under one set of instructions could not under a different set of instructions (Byer and Abrams, 1953). This hinted at an issue with using only the proportion correct and chance probability in the analysis of sensory test results. At the very least, it showed it is not a good method to compare results from different testing procedures (Ennis, 1993).

The explanation of the "paradox" is that subjects use different cognitive strategies on different testing procedures (Frijters, 1979). The 3-AFC uses a skimming strategy, whereas the triangle uses a comparison of distances strategy (Frijters, 1979). There are situations where both tests will get the same answer (O'Mahony et al., 1994). There are other situations where one test will be correct and the other will not (O'Mahony et al., 1994). Figure 2 demonstrates these four scenarios. In Figure 2A, neither strategy results in a correct response. In Figure 2B, both strategies result in a correct response. In Figure 2C, only the comparison of distances strategy results in a correct response. In figure 2D, only the skimming strategy results in a correct response. The scenario in Figure 2D is much more likely than the scenario in Figure 2C (O'Mahony et al., 1994). This is easily visualized because the lower the sensory experience (represented by either a circle or a square) on the normal distribution, the lower its likelihood of occurring. MacRae (1995b) provides a three dimensional visualization of this. This explains why subjects are expected to have a higher proportion correct for the same stimuli or products when using the 3-AFC than using the triangle (MacRae, 1995b). However, this trend will not be seen in all situations (O'Mahony et al., 1994).

Delwiche and O'Mahony (1996) compared performance on the directional and non-directional tetrad tests. In the tetrad, four samples (A, A', B, B') are given.

Subjects are told to group them into A/A' and B/B'. In the directional tetrad, subjects are told how A and B differ. In the non-directional tetrad, subjects are only told that A and B are different. It is expected that the directional tetrad will elicit the skimming strategy, where the non-directional tetrad will elicit the comparison of distances strategy (O'Mahony et al., 1994). However, this is not expected to result in the directional tetrad having superior performance to the non-directional tetrad (O'Mahony et al., 1994). This has also been shown experimentally (Delwiche and O'Mahony, 1996). Interestingly, performance on both tetrads fell below the 3-AFC, and above the triangle (Delwiche and O'Mahony, 1996).

Methods of analysis have been developed to account for the differences in tests due to cognitive strategies. These measures, d' (d-Prime) and the R-index, are meant to give a measurement scale to compare the absolute degree of sensory difference regardless of testing procedure.

1.4.4 d-Prime

The true difference between two stimuli can be visualized as the distance between the means of their normal distributions along a sensory dimension (Figure 1). This distance is referred to as d' and is measurable. Equation 5 allows calculation of d' for tests where the hits and false alarms can be measured (Lawless and Heymann, 1998). Hits are described as correctly identifying the signal sample as the signal sample (Lawless and Heymann, 1998d). Methods where this equation can be used include the A-not-A and same/different tests. For tests such as the triangle, duo-trio, and AFC's, a simple calculation from the raw data is not possible. Psychometric functions relating the proportion of correct responses observed to d' have been determined for the tests assuming perfect adherence to the hypothetical cognitive strategies discussed in 1.3.1 (Frijters, 1982; David and Trivedi, 1962; Ennis, 1993; Ura, 1960; Bradley, 1963) (Equations 6-8). David and Trivedi (1962) provide a

thorough statistical derivation of Equations 6 and 7 for the triangle and duo-trio tests, respectively. Hacker and Ratcliff (1979) derive Equation 8 for the n-AFC test.

For the purposes of simplifying these already complicated functions, several assumptions are made (Frijters, 1988). Both of the samples are assumed to produce normal distributions of sensory stimulation that are of equal variance (Frijters, 1988). Each of the functions is assumed to be independent of one another (Frijters, 1988). Due to the complexity of the functions, multiple tables have been published relating the proportion of correct answers to d' (Frijters, 1982; David and Trivedi, 1962; Ennis, 1993). Ennis (1993) currently provides the most complete tables.

$$d' = Z(hits) - Z(false\ alarms) \tag{5}$$

$$f(\delta) = 2\int_0^{\infty} \{\Phi[-z\sqrt{3} + \delta\sqrt{(2/3)}] + \Phi[-z\sqrt{3} + \delta\sqrt{(2/3)}]\}\ \partial\Phi(z) \tag{6}$$

$$f(\delta) = 1 - \Phi(\delta/\sqrt{2}) - \Phi(\delta/\sqrt{6}) + \Phi(\delta/\sqrt{2})\Phi(\delta/\sqrt{6}) \tag{7}$$

$$f(\delta) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp[-(z-\delta)^2/2]\Phi(z)^{n-1}\partial z \tag{8}$$

where d' = momentary estimate of δ

$f(\delta)$ = the function describing δ as z changes

Φ (.) = standard normal distribution function of (.)

δ = distance between the means of distributions of sensory magnitudes (Figure 1)

z = Z-score

n = number of samples in the n-AFC

From these tables, it is simple to determine the d' of duo trio, triangle and AFC tests (Ennis, 1993). Additional tables have been produced for the variance of this value (Bi et al., 1997). Bi et al. (1997) explain how to perform statistical tests to compare d' values in addition to supplying tables with recommended subject sizes.

The presence of all these tables makes d' a simple and useful scale for comparing the size of the perceived differences from different testing procedures.

Statistical analysis of d' is allowed only with an estimate of variance (Bi et al., 1997). The charts produced by Bi et al. (1997) provide a method for calculating this variance. O'Mahony and Rousseau (2002) provide an example illustrating the importance of this variance. To detect a difference with a d' of one at an α-level of 5% and a β-level of 20%, the duo-trio will require 225 subjects, the triangle 198, the 2-AFC 21, and the 3-AFC 15 (O'Mahony and Rousseau, 2002; Bi et al., 1997). The largest difference of power is seen between the directional (2-AFC, 3-AFC) and non-directional (triangle, duo-trio) tests. This is a major liability in the industrial acceptance of d' where the triangle test is commonly used and using 200 subjects is undesirable.

It must be noted, however, that d' is still not a perfect measure of sensory difference. Testing against chance missed the subtleties of cognitive strategy, which d' did not. However, the same subjects with the same stimuli can still yield different d's when pre-test (training, warm-up, preview, etc.) or in-test (repetition, instructions, etc.) conditions are varied (Dacremont and Sauvageot, 1997; O'Mahony, 1995; O'Mahony and Rousseau, 2002). Also, the theoretical basis for d' assumes change across only one sensory dimension. Ennis and Mullen (1985) show that, as the complexity of a food system increases, d' is expected to decrease for the same perceptual difference (i.e. changing from 9% sucrose to 10% is expected to elicit a higher d' in water than in a flavored drink mix).

Additionally, for d' to be an accurate estimate of sensory difference, it must be based on the cognitive strategy actually used by judges (Hautus and Irwin, 1995). Studies where the same d' is produced by different procedures have been used to support the cognitive strategies currently assumed (Kim et al., 2006; Frijters, 1979).

Both Frijters (1979) and Kim et al. (2006) stated that the lack of difference between the d's confirmed that the d's were the same. This kind of statement (lack of significant difference = no difference) only hold when power is taken into account. Neither study calculated their power level (Kim et al., 2006; Frijters, 1979).

It is probable that the cognitive strategies assumed for the calculation of d' are accurate (comparison of distances for triangle / duo-trio, skimming for n-AFC's). However, as discussed in 1.3.1, it is possible that the β-strategy may be used (Table 1) (Rousseau, 2001). The β-strategy is more powerful than the comparison of distances strategy (Rousseau, 2001). If this strategy is being used, d' values calculated for comparison of distances tests will be inflated. The R-index is an alternative scale to d' that has been used in the sensory literature (O'Mahony et al., 1985; Rousseau, 2006; Bi, 2006a; O'Mahony, 1979; O'Mahony and Goldstein, 1986; O'Mahony and Odbert, 1985). It uses a variation of the A-not-A or same-different testing procedure with a confidence scale (O'Mahony, 1979). This confidence scale expands the possible responses from two (A / Not A or Same / Different) to four (or more) possible responses (Definitely A / Maybe A / Maybe Not A / Definitely Not A or Definitely Same / Maybe Same / Maybe Different / Definitively Different). The result is a non-parametric substitute for d'. That is, R-index does not imply a specific distribution (i.e. normal) as d' does (O'Mahony, 1979).

1.5 MODIFICATIONS TO DISCRIMINATION TESTS

Several procedural modifications have been suggested for addition to discrimination testing procedures. The three main modifications are: inclusion of a warm-up, intrasubject replications, and inclusion of a 'no difference' option.

1.5.1 WARM-UP

In food sensory literature, the term 'Warm-up' refers to a "rapid increase in performance during the first few trials of a difference test" (Mata-Garcia et al., 2007).

This process occurs to various degrees depending on the degree of difference being tested, the experience of the subject with the testing procedure, as well as the experience of the subject with the stimuli being tested (Dacremont et al., 2000). It appears as though warm-up functions by drawing attention to the difference being examined (Dacremont et al., 2000). Therefore, extremely naïve subjects (unfamiliar with both the test and the stimuli) do not demonstrate the warm-up effect to a large degree because they do not know what to attend to during the warm-up (Dacremont et al., 2000). It may be possible for this group to demonstrate the warm-up effect if enough trials were presented (Dacremont et al., 2000).

Extremely experienced subjects (familiar with both the test and the stimuli) may not demonstrate the warm-up effect because they are already maximally attenuated to the test (Dacremont et al., 2000). Subjects familiar with the testing procedure, but not the stimuli (as would be the case in testing facilities with a limited subject pool) demonstrate the warm-up effect to the greatest extent (Dacremont et al., 2000). Dacremont et al., (2000) suggest that including a warm-up procedure might reduce the difference in sensitivity between experienced and inexperienced subjects.

Warm-up procedures have been examined in detail. Typically, the procedure requires no fewer than three with a maximum of seven presentations of pairs of both samples being tested (Mata-Garcia et al., 2007). A presentation of fewer than three pairs is typically referred to as a 'preview sample' (Mata-Garcia et al., 2007). This preview sample has been shown to stabilize the τ-criteria of the subjects (Mata-Garcia et al., 2007). The τ-criterion is the magnitude of the difference that the subject requires before deeming two samples different (Mata-Garcia et al., 2007). This stabilization would be expected to increase power by reducing variability, though not necessarily by increasing sensitivity (Mata-Garcia et al., 2007). The warm-up procedure (as opposed to a preview) has been shown to increase power in both ways

(Angulo et al., 2007; Mata-Garcia et al., 2007). Presenting only one sample of the pair (e.g. the reference in the duo-trio as suggested by Peryam and Schwartz (1950)) would not be expected to stabilize the τ-criteria of the subjects. A presentation of this type could, however, reduce the difference between subjects' noise distributions.

Though increased power is generally desired in a sensory test, increased sensitivity of subjects is not always a good thing. Difference tests are often used in consumer testing. In these cases, the sensitivity of test must reflect the sensitivity of the consumers, not simply be as high as possible. Warm-up procedures in this case would increase the likelihood of a Type I (false-positive) result (Mata-Garcia et al., 2007). Situations exist, however, where the consumers are more sensitive than the general population (brand loyal consumers for example). In this case it might be desired to make the subjects more sensitive if actual consumers cannot be located.

1.5.2 REPLICATIONS

One of the greatest costs associated with performing a sensory test is the recruitment of subjects (Lawless and Heymann, 1998f). It therefore makes sense to test subjects multiple times while they are already in the testing booth (Lawless and Heymann, 1998f). This way more data can be produced for the same cost. The risk of performing replications is that subjects will perform differently as they are given more replications. That is not to say that subjects are expected to get all replications correct or incorrect, but that subjects may become fatigued or warmed-up. Dacremont and Sauvageot (1997) examined the effect of replications on various liquid foods (soda, milk, fruit syrup). They did not find any consistent change in performance over four replications, though they still suggested that an increase in performance might be seen due to warm-up effects (Dacremont and Sauvageot, 1997). Pfaffmann (1953) found that fatigue should not be assumed on the basis of a high numbers of samples, as long as proper sensory practices are followed. The degree of fatigue is dependent on the

type of food being examined (hot sauce is more fatiguing than soda) (Pfaffmann, 1953). If there is significant fatigue or undesired warm-up effects, the replications could be thrown out, and the first test for each subject could be analyzed as if no replicates had been conducted.

Interpretation of replicated data is not always straightforward. The most cost efficient scenario is to test subjects multiple times and count each repetition as if the data were from another individual subject (Lawless and Heymann, 1998c). This would reduce the number of subjects needed, saving money. Most statistical methods assume that every data point is independent from every other data point, that one rating does not impact the other, replications have the potential to violate this assumption (Lawless and Heymann, 1998c). Consider coin tosses (Angulo et al., 2007). Tossing one coin ten times is expected to produce equivalent results as tossing ten coins once. This is because the differences between coins are small. Human subjects, however, have large differences between them. In order consider one subject's ten ratings as equivalent to ten subjects' one rating; the variation between subjects must be comparable to the variation within one subject's multiple ratings.

Several different methods have been proposed to deal with replications in discrimination tests (Smith, 1981; Priso et al., 1994; Ennis and Bi, 1998; Dacremont and Sauvageot, 1997; Kunert and Meyners, 1999; Brockhoff and Schlich, 1998). Equations 9.1-9.3 discuss the methods for dealing with replicates from most to least conservative. N_{eq} is the "equivalent panel size," that is the number of subjects without replication that would be required for the same power level. The most conservative method is to count each subject (N_j) only once, regardless of the number of replications (n_R) (8.1). The least conservative method is to assume that each individual test is equivalent to a unique subject (8.3). The third method is to use an intermediate value by accounting for the differences between inter and intrasubject

variations by dividing by some factor ($X_{dispersion}$) where $X_{dispersion}$ falls between 1 and n_R.

$$N_{eq} = N_J \tag{9.1}$$

$$N_{eq} = \frac{N_J n_R}{X_{dispersion}} \tag{9.2}$$

$$N_{eq} = N_J n_R \tag{9.3}$$

The most conservative method is to consider the answers to all replications as one data point (Equation 9.1). This method still effectively increases the power by the reduction of standard error, due to an increase in possible data points. Without replication, subjects can either give a correct (1) or an incorrect (0) answer. With replications (four for example) subjects can produce a wider variety of data points (0/4, 1/4, 2/4, 3/4, or 4/4). This has an impact on the standard error of the test. Imagine the comparison of two 100-subject tests, one with no replications, one with four replications. Let the proportion correct be 3/4 for both tests. With no replications, this means that 25 subjects were incorrect and 75 were correct. This results in a mean of 0.75 with a standard error of 0.0435. With four replicates, 0.0435 represents the maximum standard error possible (25 = 0/4, 75 = 4/4). Assuming responses are distributed around the mean (33 = 2/4, 34 = 3/4, 33 = 4/4) the standard error is halved (0.0204). This effectively increases the power of the test even though the subject size has not changed. The benefit of this method is that fewer assumptions must be made compared to the alternative methods of dealing with replications. The data are analyzed statistically using one of the methods discussed earlier, comparing the proportion correct to chance or calculating d'. Theoretically, individual subjects' d's could be calculated. However, if a portion of subjects performs below chance (which is expected), their d's will not be easily calculated using available charts and/or software.

Smith (1981) proposed a method for determining if replications can be pooled and counted as individual subjects. Consider an experiment with two replications. There will be four kinds of subjects. Those who get both tests wrong, those who get both tests correct, those who get only the first correct (r_1), and those who get only the second correct (r_2). Smith's method (Smith, 1981) proposes that if r_1 and r_2 are significantly different, replicates cannot be pooled. If they are not significantly different, replicates can be pooled. This method does not allow for partial pooling of replications, and is increasingly complicated as the number of replications increases.

The β-binomial method accounts for the difference between inter and intra subject variations and allows for a proportional increase in N_{eq} (Ennis and Bi, 1998). Several parameters have been used for this purpose (Priso et al., 1994; Brockhoff and Schlich, 1998; Ennis and Bi, 1998). One such parameter, γ, is frequently discussed in the sensory literature (Angulo et al., 2007; Ennis and Bi, 1998; Bi and Ennis, 1999a; Bi and Ennis, 1999b; Liggett and Delwiche, 2005) and will be discussed further here.

Theoretically γ can be range from negative to positive one. γ values greater than zero indicate overdispersion. Overdispersion is when intersubject variation is greater than intrasubject variation (Ennis and Bi, 1998). Under these conditions, subject replications cannot be considered equivalent to separate subjects (Ennis and Bi, 1998). Underdispersion exists theoretically, but is rarely encountered. For underdispersion to occur, intersubject variation must be lower than intrasubject variation. Imagine ten rolls of ten six-sided dice. Underdispersion would be if each individual die produced the same ten digit random number. Because of underdispersion's rarity, γ is traditionally considered to range from zero to one. Data with a γ not significantly greater than zero are then considered to have "no overdispersion" (Ennis and Bi, 1998). Under these conditions, an experimenter could justify counting each replication as a unique subject (Ennis and Bi, 1998). When

gamma is significantly larger than zero, replications can still be used to gain power. However, as gamma increases, the power gained by each replicate is reduced. Equation 10 can be used to calculate the "equivalent panel size," a factor that describes how many individual subjects would be needed to have the same power as the replicated data (Liggett, and Delwiche, 2005). The "equivalent panel size" approaches the number of subjects as gamma increases.

$$N_{eq} = \frac{N_J n_R}{1 + (n_R - 1)(\gamma)} \tag{10}$$

where N_J = Number of Subjects

n_R = Number of Replicates

Liggett and Delwiche (2005) noted that the degree of overdispersion was not consistent over time. They also noted that, although the duo-trio had higher overdispersion than other methods, it is difficult to predict the relative degree of overdispersion based on the testing method used (Liggett and Delwiche, 2005). This suggests that γ should always be calculated before the β-binomial method is used on data with replicates (Liggett and Delwiche, 2005). Equations 11.1-11.4 allow the calculation of γ (Ennis and Bi, 1998). It is possible, however, to determine the goodness of fit of the binomial distribution without estimating γ. Tarone's Z-statistic (Equations 12.1-12.3) (Ennis and Bi, 1998), as determined by Tarone (1979) allows for this. The null hypothesis is that the underlying distribution is Binomial. The alternative hypothesis is that the underlying distribution is β-binomial.

Bi and Ennis (1999a) provide equations and tables for calculating the power level of discrimination tests using replicates. Bi and Ennis (1999b) provide tables for determining the number of correct assessments needed for significance at various levels of α, β, and γ for several testing procedures. Kunert and Meyners (1999) suggest that a well-designed experiment should always have no overdispersion.

However, due to the relative ease of γ's calculation with computer programs such as Excel, this assumption is not necessary. It has been noted that, regardless of the statistical justification, counting ten replications as ten subjects does not produce as representative of a sample as ten subjects (Angulo et al., 2007).

$$\gamma = \frac{nS}{[\bar{p}(1-\bar{p})\bar{k}(n-1)] - \frac{1}{n-1}} \tag{11.1}$$

$$S = \sum_{i=1}^{k} (\hat{p}_i - \bar{p})^2 \tag{11.2}$$

$$\bar{p} = \sum_{i=1}^{k} \hat{p}_i / k \tag{11.3}$$

$$\hat{p}_i = x_i / n, i = 1,2,...k \tag{11.4}$$

where γ = degree of overdispersion

n = number of replications

S = variance

$\bar{p}$ = average proportion correct

$\bar{k}$ = number of subjects

$\hat{p}_i$ = proportion correct for the ith subject

x_i = number of correct responses for the ith subject

$$Z = \frac{E - nk}{\sqrt{2kn(n-1)}} \tag{12.1}$$

$$E = \sum_{i=1}^{k} \frac{(x_i - n\hat{p})^2}{\hat{p}(1-\hat{p})} \tag{12.2}$$

$$\hat{p} = \sum_{i=1}^{k} \frac{x_i}{nk} \tag{12.3}$$

where Z = Tarone's Z-statistic

n = number of replications

k = number of subjects

$\hat{p}$ = average proportion correct

x_i = number of correct responses for the ith subject

1.5.3 "NO DIFFERENCE" OPTION

A third modification often discussed in the sensory literature is the inclusion of a "no difference" option in the test procedure. The idea is that subjects who cannot tell a difference will no longer guess, instead choosing the "no difference" option. Braun et al. (2004) examined d's for 2-AFC and 2-AC (a 2-AFC with a "no difference" option) procedures. The 2-AC is thought to employ a different cognitive strategy than the 2-AFC and therefore different psychometric function is used (Braun et al., 2004). Gridgeman (1959) discussed the statistics of the 2-AC, determining that it is more powerful than the 2-AFC. It was warned that, in practical use, this might not always be the case (Gridgeman, 1959).

An additional complication is that all of the statistical methods discussed in this paper assume a forced-choice test. That is, when subjects cannot detect a difference they are forced to guess between the available options. By providing a "no difference" option this can no longer be assumed, complicating analyses. This is especially when one considers the response bias associated with this options. The 2-

AFC does not employ a τ-criterion (Braun et al., 2004). In the 2-AC, Subjects must decide if the difference is big enough to be called a difference (a τ-criterion), before they choose which stimuli to select (Braun et al., 2004). Because of these complications and the relative lack of benefits, "no difference" options are not frequently used (Braun et al., 2004).

1.6 CHOOSING A DISCRIMINATION TEST

The previous sections have illustrated the differences in available discrimination testing methods, as well as methods for their analysis. Despite the fact that statistics such as d' can be used to compare results from multiple tests, these tests are still not interchangeable. Though it is expected that the same subjects with the same stimuli will produce the same d' on the 2-AFC, 3-AFC, triangle, duo trio, etc., the power of these tests is still different (Ennis, 1993; Bi et al., 1997). This power difference is exposed in the statistical analyses available for d' values discussed in 1.5.2 (Ennis, 1993; Bi et al., 1997).

Due to their increased power, directional tests are preferred to non-directional tests. Table 2 illustrates how many more subjects are required in the triangle test for the same level of power in the 2-AFC (Ennis, 1993). However, directional tests can only be used in situations where the difference is considered easy to explain with a simple term (Lawless and Heymann, 1998c). This term may be a magnitude term such as 'strongest' or a descriptive term such as 'sweetest.' It is generally assumed that fluent English speaking panelists will understand these types of basic terms. However, confusion does occur between seemingly simple terms. Sour/bitter confusion is a common example (O'Mahony et al., 1979). Subjects without a clear understanding of the term being used are likely to perform poorer than those that understand the term clearly. Often, the difference is too complex to describe with one

Table 2. Number of subjects required to detect a difference of size δ with an α-risk of 5% and a β-risk of 20% (A) and 10% (B) using either the Triangle or 2-AFC procedures (from Ennis, 1993). The ratio of subjects required for the Triangle versus the 2-AFC to maintain equal power is also provided (i.e. to detect a difference of $\delta = 0.5$ at equal α-risk and β-risk it takes 35 times more subjects for the Triangle than for the 2-AFC).

A. α-risk = 5% / β-risk = 20%

δ	**Triangle**	**2-AFC**	**Triangle : 2-AFC**
0.5	2742	78	35 : 1
1	197	20	10 : 1
1.5	47	9	5 : 1
2	19	6	3 : 1

B. α-risk = 5% / β-risk = 10%

δ	**Triangle**	**2-AFC**	**Triangle : 2-AFC**
0.5	3810	108	35 : 1
1	276	27	10 : 1
1.5	66	12	6 : 1
2	26	7	4 : 1

term, or consensus over the difference cannot be reached at bench-top. In these situations, non-directional tests are usually used (Lawless and Heymann, 1998c).

However, the desire for increased power often necessitates the use of directional tests when conditions are not ideal. The simplest, though least scientific method, is to assume that the term chosen will represent the actual difference to the subjects in the test. Definitions may be provided with the purpose of explaining the term(s) used (Giboreau et al., 2007). Giboreau et al. (2007) attempt to provide guidelines for those wishing to make definitions. Another option is providing subjects with reference samples to illustrate the type of difference (Rainey, 1986). This is a key feature in training a panel (Rainey, 1986). Training, by definition, is an undesired feature when conducting an untrained panel (O'Mahony and Rousseau, 2002). A similar option would be to provide a warm-up (Mata-Garcia et al., 2007). As discussed in 1.5.1, a warm-up procedure might increase the sensitivity of subjects beyond the level of the population they are meant to represent. Methods for using directional tests in non-directional situations are discussed in the next section.

1.7 NOVEL DISCRIMINATION TESTS

It has been suggested that a modification of an AFC procedure might make it an alternative to the triangle test. The 2-AFC was used as the base test instead of the 3-AFC for several reasons. Most simply, the 2-AFC has a longer publishing history than the 3-AFC, and is more popular. Because of this, a procedure based on the 2-AFC then has greater potential for quick acceptance into the sensory world. Debate exists as to whether the 2-AFC or 3-AFC is more powerful. The tables provided by Bi et al. (1997) show that fewer subjects are required with the 3-AFC to find the same level of difference as the 2-AFC. Sequential Sensitivity Analysis (SSA), however, predicts that the 2-AFC will be more powerful (Dessirier and O'Mahony, 1999).

SSA examines testing procedures in the light of the comparisons of strong (S) and weak (W) samples they produce (Dessirier and O'Mahony, 1999). The possible orders, in order of increasing difficulty of discrimination, are: WS, SW, WW, and SS. The 2-AFC only contains WS and SW comparisons, whereas the 3-AFC contains all of them. A WWS order in the 3-AFC would be considered a WW and a WS comparison in SSA. Additional practical benefits of the 2-AFC include its speed compared to the 3-AFC. The 2-AFC uses fewer samples, reducing the risk of adaptation and fatigue. These benefits as well as SSA have been used to justify the use of the 2-AFC over the 3-AFC (Dessirier and O'Mahony, 1999). Conveniently, however, the novel discrimination tests discussed can be easily converted to 3-AFC-based tests if evidence supports its superiority.

Theime and O'Mahony (1990) described a testing procedure where subjects were given the two samples to be discriminated between three and ten times and were told to define the difference. This definition contains both the quality (sweetness, for example) and direction of the difference (i.e., A is sweeter than B). This attribute was then used in a traditional 2-AFC test. A similar procedure was mentioned by Pfaffmann (1953). This procedure was described as a "warmed-up paired comparison" (Theime and O'Mahony, 1990). The authors warned that the warm-up process might make the subjects sensitive to the point where their results would not be applicable in an untrained panel (Theime and O'Mahony, 1990). This increase in sensitivity may be caused by the warm-up itself or some effect of the subjects defining the difference themselves (O'Mahony et al., 1988).

Theime and O'Mahony (1990) compared this test to the duo-trio using NaCl solutions. Subjects performed a session of fourteen constant reference duo-trio's with the NaCl as the standard followed by a session of fourteen of one of several other tests. Judges performing the "warmed-up paired comparison" in the second session

got significantly more tests correct ($p \leq 0.0007$) than in the first session. Subjects performing an A-Not A w/ warm up, duo-trio with water standard, and duo-trio with water standard and warm-up all performed significantly better as well, but had a smaller increase than the "warmed-up paired comparison". No significant increases in performance were seen in judges performing the duo-trio or A-Not A in the second session. The "warmed up paired comparison" was not directly compared to the paired comparison.

O'Mahony et al. (1988) found similar results using triangle tests. Adding a warm-up increased the number of correct responses using NaCl solutions as well as orange juice ($p = 0.001$ and $p < 0.0005$ respectively) versus not including a warm-up. Allowing subjects to define the difference they were looking for significantly improved performance testing NaCl solutions ($p = 0.0068$). Others have used the "warmed-up paired comparison" in situations where increased sensitivity of subjects was desired (Delwiche et al., 2001; Jiamyangyuen et al., 2002). Delwiche et al. (2001) used the "warmed-up paired comparison" to examine solutions of several pure bitter compounds. Jiamyangyuen et al. (2002) used the same method to examine the impact of different wooden sticks on unfrozen ice cream mix. In both these cases, the authors stated a desire for very high sensitivity with limited training, not results representative of the general population (Delwiche et al., 2001; Jiamyangyuen et al., 2002).

A modification to the "warmed-up paired comparison" procedure would be to have subjects receive the samples only once before defining their attribute. This would be defined as a preview (Mata-Garcia et al., 2007) instead of a complete warm-up. It was shown by Rousseau et al. (1999) that preview samples (called familiarization samples) function differently than full warm up procedures. An increase in d' was seen due to the preview samples (Rousseau et al., 1999). However,

it was attributed to a stabilization of tau criteria, rather than an increase in subject sensitivity (Rousseau et al., 1999). This suggests (though must still be confirmed) that a preview sample can be used without the increased sensitivity resulting from warm ups. Potentially, this would allow a 2-AFC procedure to be used in situations where the nature of the difference is not known or easily defined. This procedure could be defined as a Subject-Defined 2-AFC or SD-2-AFC. While this would reduce the likelihood of a warm-up effect, having subjects define the difference incorporates other complexities to the procedure as well.

The attributes used to describe the difference in an extremely simple system (solutions of NaCl at two different concentrations, for example) would be expected to be relatively consistent across subjects (i.e., most subjects would be expected to say one sample is saltier than the other). The attributes used to describe the difference in an extremely complex system would be expected to be less consistent across subjects. When testing the impact of different woods on ice cream, Jiamyangyuen et al. (2002), noted a wide range of subject defined attributes: "burnt, creamy, cucumber, dry wood, fresh wood, green, oily, brown paper, rancid, sweet, and wet biscuit." Subjects were able to use these attributes to accurately discriminate ice cream mixes treated with different woods (Jiamyangyuen et al., 2002). This suggests that, while most of the literature describes subject-defined attributes in simple systems, it may still be effective in more complicated systems.

1.8 SUMMARY OF CHAPTER 1

A wide variety of sensory discrimination procedures and methods of analysis have been discussed. The goal of all of these procedures and analyses is to determine, in the most reliable and efficient way possible, if two products are different in a meaningful way. Directional tests have been shown to be more likely to find statistically significant differences at the same level of difference (i.e. the same d')

than non-directional tests (Figure 2) (O'Mahony et al., 1994). This has been attributed to the different cognitive strategies elicited by directional and non-directional tests. The use of directional procedures, however, is only possible when the nature of the difference is known and easily described. This is not always the case.

Alternatively, the addition of a warm-up procedure before a discrimination test has been shown to increase the d' found for two products compared to the same discrimination test without a warm-up procedure (Mata-Garcia et al., 2007). This effect has been attributed to subjects learning the magnitude and type of difference to focus on. This can be used either on directional or non-directional procedures. The possibility exists, however, that the addition of a warm-up procedure functions as training. The warm-up would then make the panel more sensitive than the population of interest (Mata-Garcia et al., 2007). This increases the chance of finding a statistically significant difference that is not important to the target population, a highly undesirable result.

The SD-2-AFC has been proposed as a method of using a directional test where the nature of the difference is not known or easily described. In order for the SD-2-AFC to be successful and useful for discriminating samples in situations where the 2-AFC cannot be used, several conditions must be met. First of all, the details of the procedure must be described clearly. This will avoid confusion between the SD-2-AFC and tests such as the "warmed up paired comparison" which are not suitable for the desired used. Secondly, for the SD-2-AFC to be useful, it must be more powerful than the Triangle test. If this condition were not met, there would be little reason to argue for using the SD-2-AFC instead of the Triangle in industry.

The possibility exists as well, that the SD-2-AFC could perform better than the traditional 2-AFC. While this is not expected, it bears examination. Additionally, it is unclear if the required preview sample would increase the chance of finding a statistically significant difference that is not important to the target population. This needs to be determined before the SD-2-AFC can be considered a viable testing procedure. Finally, the SD-2-AFC must be examined over a wide variety of product complexities to determine how varied the attributes selected by subjects become.

CHAPTER 2

OBJECTIVES AND HYPOTHESES

2.1 OBJECTIVES

This research has the following objectives:

1. Define the protocol for a new discrimination test: the Subject Defined 2-Alternative Forced Choice (SD-2-AFC).
2. Determine if the SD-2-AFC will perform at a greater or equal sensitivity to the Triangle test where the traditional 2-AFC is not suitable.
3. Determine the efficacy of the SD-2-AFC in situations where the traditional 2-AFC would be used (i.e., simple changes).
4. Determine the efficacy of the SD-2-AFC in situations where the traditional Triangle test would be used (i.e., complex changes).
5. Explain the differences between the SD-2-AFC, 2-AFC and Triangle tests.
6. Determine the effect of preview samples on panel sensitivity.

2.2 HYPOTHESES

This research has the following hypotheses:

1. The SD-2-AFC will not be a suitable alternative for the Triangle test where the 2-AFC is not suitable.
2. The SD-2-AFC will perform as well as the 2-AFC in situations where the 2-AFC would be used.
3. The SD-2-AFC will perform significantly worse than the triangle test in situations where the triangle test would be used.
4. The preview samples alone will not significantly increase sensitivity.

2.3 OVERALL EXPERIMENTAL PLAN

The Warmed-Up Paired Comparison was modified so that a preview was used instead of a full warm up procedure. This was done to minimize the differences

between the subjects and completely naïve subjects that would normally be used in discrimination testing. This new procedure was named the Subject Defined 2-AFC (SD-2-AFC). The SD-2-AFC was tested over four products of differing complexities. This allowed analyses to examine the effects of complexity (i.e., a simple system of differing sweetness in Kool-Aid verses a more complex system of differing Monosodium Glutamate in Soup) on the SD-2-AFC's performance. The magnitude of difference between stimuli in each product was varied as well, from very obvious to very difficult (though still discriminable). This allowed analyses to examine the effects of difference magnitude (i.e. d') on the SD-2-AFC's performance. Conventional testing procedures (2-AFC and Triangle) were tested simultaneously, with and without a preview condition. This allowed analyses to 1) directly compare the SD-2-AFC to its alternatives and 2) determine if the addition of a preview condition made the panel significantly different than a naïve panel.

CHAPTER 3

MATERIALS AND METHODS

3.1 SUBJECTS

606 subjects between the ages of 18 and 65 were recruited from the Cornell University Campus over the course of several months. Informed consent (Appendix I) was obtained and subjects were either entered in a raffle or received $2 as a token incentive. Table 3 summarizes the number of subjects in each testing condition.

3.2 MATERIALS

Four different food products were used to simulate four product development scenarios. Scenarios were divided into simple and complex stimuli. Simple stimuli were those where the change was expected to modify a minimum of sensory dimensions, sweetness and sourness. Complex stimuli were those where the change was expected to modify several sensory dimensions and the dominant dimension could not easily be predicted.

3.2.1 SIMPLE STIMULI

In the first simple stimuli condition, grape flavored Kool-Aid brand unsweetened drink mix from the same lot was used. Solutions were made at the manufacturer's recommended concentration (2.1g Mix / 1.0 L Solution) with 9% or 10% (w/v) sucrose added. 10 milliliters of a given solution was presented in lidded two ounce (59 mL) plastic sample cups.

In the second simple stimuli condition, orange flavored Tang drink mix from the same lot was used. Solutions were made at the manufacturer's recommended concentration (4.65% w/v) with 0.0 or 3.0 g/L citric acid monohydrate added. 30 milliliters of a given solution was presented in lidded two ounce (59 mL) plastic sample cups.

Table 3. Number of subjects in each test used. All subjects within a given stimulus are unique

	Kool-Aid	**Tang**	**Tea**	**Soup**
Triangle	39	42	34	33
Triangle w/ Preview	40	42	34	31
SD-2-AFC	40	41	33	32
2-AFC	41	42	N/A	N/A
2-AFC w/ Preview	40	42	N/A	N/A

3.2.2 COMPLEX STIMULI

In the first complex stimuli condition, Wegmans Brand Iced Tea Mix from the same lot was used. The principle flavorants of the mix were sucrose, citric acid, powdered tea, and lemon flavors. Solutions were 8% w/v (recommended concentration) and 12% w/v. 10 milliliters of a given solution was presented in lidded two ounce (59 mL) plastic sample cups.

In the second complex stimuli condition, Swanson's vegetarian vegetable broth was used. Solutions were served either with no modifications (control) or with 3.5g/L commercially available food-grade monosodium glutamate (MSG) added. 10 milliliters of a given solution was presented in lidded two ounce plastic (59 mL) sample cups. Since enough broth from the same lot could not be acquired, all broth used on a given day was commingled before modifications were made.

3.3 METHODS

All testing procedures used paper ballots that were tallied by hand (Appendices II and III). The methods used were based on the conventional triangle and two alternative forced choice (2-AFC) tests. The modifications are summarized in Table 4. Subjects testing simple stimuli were randomly assigned to one of the five procedures (triangle, triangle with preview, 2-AFC, 2-AFC with preview, or Subject Defined 2-AFC (SD-2-AFC)). Subjects testing complex stimuli were randomly assigned to one of three procedures (triangle, triangle with preview, or SD-2-AFC). In addition to the preview sample (if applicable) each subject performed four replicates of the procedure. Each replicate had a randomized order. A preview sample was defined as a presentation of each test stimulus (i.e. A and B) before beginning the actual testing procedure. Preview ballots were provided during the preview sample mainly to insure that subjects were attentive during tasting (Table 4). In the SD-2-

Table 4. Summary of Testing Procedures

	Preview Procedure	**Samples Presented**
Triangle	No Preview	A, B, and A' or B'
2-AFC	No Preview	A and B
Triangle with Preview	Give A, B with a statement indicating that: "Sample A is different than Sample B." The subject then indicates that they can or cannot tell that they are different	A, B, and A' or B'
2-AFC with Preview	Give A, B with a statement indicating that: "Sample A is sweeter than Sample B." The subject then indicates that they can or cannot tell that A is sweeter than B	A and B
Subject Defined 2-AFC (SD-2-AFC)	Give A, B with a statement indicating that: "Sample A is different than Sample B." The subject then indicates how they think they are different in the format "A is [blank] than B"	A and B

AFC, the preview ballot also served as the opportunity for the subject to provide the attribute word(s) used in the subsequent test.

3.4 ANALYSIS

The four replicates that each subject performed were summed to give each subject a possible score between zero and four. This number was considered to be one data point; that is to say, each subject was counted as one subject (not four). T-Tests were conducted to determine if subjects were able to discriminate the samples within a condition. Chance levels were 1.33 for the Triangle's and 2.00 for both the 2-AFC's and SD-2-AFC's. While a range of differences was desired, it was important that subjects could significantly discriminate all the samples in order to perform further analyses. Values of d' for each condition were then obtained from Ennis (1993). Variances of the d' values were obtained from Bi and Ennis (1997). Chi-Squares were calculated on the d' values within each stimulus to determine the relative performance of the testing procedures under the different conditions. At this point, it was determined if the SD-2-AFC appeared to be a viable alternative for the 2-AFC or triangle tests. Further analyses were performed to provide an explanation for the performance of the SD-2-AFC.

First, the effect of the preview condition was examined. For procedures where a preview was given, subjects were divided into two groups by their responses and the same analysis as above was performed to look for differences between groups. Subjects indicating they could detect a difference in one group, subjects indicating they could not detect a difference in another. T-tests were run on the d's of these groups to determine if the preview sample had a positive, negative, or inconsistent effect on performance within each condition. For the SD-2-AFC condition, an analysis was developed to determine how the attributes selected by subjects affected to their performance. Subjects were grouped based on the attributes they selected.

Section 4.2.1 explicitly describes how subjects were grouped for each product. T-tests were then run on the d's of the groups within each condition to determine if there was a significant difference between the performance of the groups. Finally, effects due to presentation order (i.e. A A' B versus A B A') as well as replicate order (i.e. Replicate 1 versus Replicate 4) were examined as well to confirm that they were not significant.

CHAPTER 4

RESULTS

Tests were first analyzed to insure that the food samples could be significantly discriminated. Tables 3 and 5 summarize the number of subjects, means, and standard errors of all conditions. Table 6 summarizes the p-values for the t-tests against chance for all conditions. All samples were significantly discriminated at the 0.10 α-level. With the exception of the Kool-Aid Triangle w/ Preview all samples were significantly discriminated at the 0.05 α-level. This demonstrates that the tests were capable of detecting differences under the conditions of the study and justifies further analysis. The fact that the one condition could not be discriminated at the 0.05 α-level suggests that the difference between the samples was as close as could be used without requiring a larger number of subjects. This is important because a wide range of sensory differences was desired.

4.1 PROCEDURE EFFECTS

d's were calculated for each testing condition using the tables by Ennis (1993) (Table 7). χ^2 tests were run on the d's within each product using the estimate of variance provided by Bi et al. (1997). A significant difference was found between the Tea conditions. The SD-2-AFC performed significantly worse than both Triangles. Appendix IV provides the calculations for the Tea conditions as a sample. More generally, within a given food product, the SD-2-AFC always performed worst (i.e. had the lowest d'). This strongly suggests that the SD-2-AFC is not superior to either the Triangle or 2-AFC.

4.2 SD-2-AFC

The SD-2-AFC deviates from the 2-AFC in two main ways, and it is likely that at least one of these is the cause of its underperformance. First, subjects are exposed to the samples before testing, as would be done in a preview sample. Second, subjects

Table 5. Mean scores with standard errors in parentheses.

	Kool-Aid	**Tang**	**Tea**	**Soup**
Triangle	1.62 (0.140)	2.67 (0.173)	3.12 (0.168)	1.85 (0.190)
Triangle w/ Preview	1.55 (0.164)	2.45 (0.178)	3.44 (0.165)	2.10 (0.193)
2-AFC	2.88 (0.149)	3.57 (0.128)	N/A	N/A
2-AFC w/ Preview	3.08 (0.158)	3.64 (0.131)	N/A	N/A
SD-2-AFC	2.83 (0.208)	3.17 (0.206)	3.58 (0.163)	2.53 (0.246)

Table 6. P-Values of T-Tests versus Chance Probability. *Only testing condition not significant at the $p < 0.05$ level.

	Chance Probability	Kool-Aid	Tang	Tea	Soup
Triangle	0.33	0.0259	<0.0001	<0.0001	0.0054
Triangle w/ Preview	0.33	0.0969*	<0.0001	<0.0001	0.0002
2-AFC	0.50	< 0.0001	<0.0001	N/A	N/A
2-AFC w/ Preview	0.50	< 0.0001	<0.0001	N/A	N/A
SD-2-AFC	0.50	0.0002	<0.0001	<0.0001	0.0193

Table 7. d-Prime scores for all testing conditions. *Different superscript letters indicate significant difference at 0.05 level.

	Kool-Aid	Tang	Tea*	Soup
Triangle	0.90	2.32	2.98^{A}	1.26
Triangle w/ Preview	0.88	2.04	3.61^{A}	1.59
2-AFC	0.82	1.75	N/A	N/A
2-AFC w/ Preview	1.03	1.90	N/A	N/A
SD-2-AFC	0.76	1.15	1.76^{B}	0.48
P-values (Chi Square)	0.979	0.162	0.0167	0.105

choose their own test attribute. The first difference was examined by adding a preview to the 2-AFC and Triangle procedures. The preview did not have a significant or consistent effect on the d' values of the Triangle or 2-AFC for any of the stimuli. This suggests that inclusion of a preview sample alone does not account for the significant underperformance of the SD-2-AFC. The attribute selection aspect of the SD-2-AFC is likely to explain the underperformance of the SD-2-AFC.

4.2.1 EFFECTS OF ATTRIBUTES SELECTION AND PREVIEW

Subjects were divided into groups based on the attribute they selected for use in the 2-AFC. Table 8a summarizes the number of subjects in each group, discussed later. For all tests, a plurality of subjects used attributes that were similar enough to be grouped. These attributes were categorized as typical. All other subjects were categorized as atypical. Table 8b summarizes the attributes used to define each group. For the simple stimuli, the attribute given on the 2-AFC was used to define the typical group (sweeter for Kool-Aid, and sourer for Tang). Increasing concentrations of sweetener can result in a perceived reduction in sourness and vice versa, due to sweet-sour mixture suppression (Pelletier et al., 2004). Because of this, 'less sour' was considered equivalent to 'sweeter' and 'less sweet' was considered equivalent to 'sourer'. For the Tea stimuli, the majority of subjects used the attribute 'sweeter,' so this was used to define the typical group. For the Soup stimuli, the majority of subjects used 'saltier' and/or 'stronger.' Five subjects described the higher MSG level as 'saltier', five as 'stronger', and five as 'saltier and stronger.' Due to the frequent overlap of these attributes both attributes were used to define the typical group. Any other responses were categorized as atypical.

The atypical group was then divided into two additional groups. Attributes that were perfect opposites of the typical attributes (less sweet instead of sweeter, for example) were categorized as inverted. All other atypical attributes were categorized

Table 8a. Number of subjects in each group used for analysis of the SD-2-AFC

	Typical	**Atypical**	
	Non-Inverted	**Inverted**	**Unique**
Kool-Aid	28	10	2
Tang	28	8	5
Tea	20	5	8
Soup	15	10	7

Table 8b. Criteria used to group subjects in analysis of the 2-AFC. The number of subjects who used a given criterion is in parentheses. *Two or fewer subjects selected the same criteria.

	Typical		**Atypical**			
	Non-Inverted		**Inverted**		**Unique**	
Kool-Aid	Sweeter (25)	Less Sour (3)	Less Sweet (8)	Sourer (2)	Other* (2)	
Tang	Sourer (21)	Less Sweet (7)	Less Sour (1)	Sweeter (7)	Other* (5)	
Tea	Sweeter (20)		More Sour (3)	Less sweet (2)	Stronger (3)	Other* (5)
Soup	Saltier (9)	Stronger (6)	Less Salty (9)	Weaker (1)	Other* (7)	

as ‘unique.’ For this analysis, the ‘typical’ group was renamed ‘non-inverted’ to emphasize its opposition to the ‘inverted’ group. Table 8b shows the attributes used by subjects and their grouping into categories.

Table 9 summarizes the d’ values of each group for all conditions. Table 10 summarizes the p-values for the t-tests comparing groups. Tables 11a-d summarize the mean and standard error for all SD-2-AFC conditions. Comparisons were made between typical/atypical groups as well as the inverted/not inverted groups. In some instances, the d’ for a group was negative and could not be calculated. In these cases a d’ of zero was used. It was assumed that a d’ significantly greater than a d’ of zero would also be significantly greater than a d’ below zero. The weakness to this approach is that there is no measure of variance for the negative d’ value because the method used to estimate variance cannot account for negative d’s.

The typical group performed significantly better than the atypical group at the 0.05 α-level for the Kool-Aid SD-2-AFC. The atypical group performed below chance, so its exact d’ could not be determined. A d’ of zero was used for statistical testing as described above. More generally, the typical group performed better (though not significantly) than the atypical group for all of the SD-2-AFC conditions (Table 9). The non-inverted group performed significantly better than the inverted group at the 0.05 α-level for the Kool-Aid and Tea condition. The inverted group in the Kool-Aid condition performed below chance, so its exact d’ could not be determined. A d’ of zero was used for statistical testing as described above. More generally, the non-inverted group performed better (though not significantly) than the inverted group for all of the SD-2-AFC conditions (Table 9). There was also a negative correlation ($R^2 = 0.937$) between the proportion of subjects inverted and the d’ for the group (Figure 8). That is to say, the larger the difference, the smaller the proportion of inverted subjects.

Table 9. d' values for the different preview groups.

		Soup	Kool-Aid	Tea	Tang
Triangle w/ Preview	Confident	1.84	1.09	3.53	2.03
	Not Confident	1.08	negative	3.94	2.1
2-AFC w/ Preview	Confident	1.09	N/A	N/A	2.07
	Not Confident	0.95	N/A	N/A	1.3
SD-2-AFC	Typical	0.74	1.24	2.51	1.45
	Atypical	0.26	negative	1.13	0.63
SD-2-AFC	Non-Inverted	0.74	1.21	2.52	1.45
	Inverted	0.09	negative	0.54	0.45
	Unique	0.51	1.62	1.62	0.95

Table 10. P-Values for difference between d's of the indicated groups. *Atypical / Not Confident / Inverted group had a negative d', so a d' of zero was used.

	Comparison	Kool-Aid	Tang	Tea	Soup
Triangle w/ Preview	Confident vs. Not Confident	0.0321*	0.475	0.3886	0.238
2-AFC w/ Preview	Confident vs. Not Confident	0.413	0.197	N/A	N/A
SD-2-AFC	Typical vs. Atypical	0.000543*	0.104	0.0662	0.229
SD-2-AFC	Non-Inverted vs. Inverted	0.000804*	0.0586	0.00750	0.123

Table 11a-d. Descriptive statistics for the different preview groups for the a) Triangle w/ Preview, b) 2-AFC w/ Preview, c) SD-2-AFC – Typical vs. Atypical, and d) SD-2-AFC Non-Inverted vs. Inverted vs. Unique

a	Confident			Not Confident		
Triangle w/ Preview	N	Mean	Standard Error	N	Mean	Standard Error
Soup	20	2.30	0.219	11	1.73	0.359
Kool-Aid	26	1.73	0.204	14	1.21	0.261
Tea	27	3.41	0.194	7	3.57	0.297
Tang	36	2.44	0.197	6	2.50	0.428

b	Confident			Not Confident		
2-AFC w/ Preview	N	Mean	Standard Error	N	Mean	Standard Error
Kool-Aid	25	3.12	0.203	15	3.00	0.258
Tang	35	3.71	0.113	7	3.29	0.565

c	Typical			Atypical		
SD-2-AFC	N	Mean	Standard Error	N	Mean	Standard Error
Soup	15	2.80	0.312	17	2.29	0.371
Kool-Aid	28	3.24	0.202	12	1.92	0.417
Tea	20	3.85	0.109	13	3.15	0.355
Tang	28	3.39	0.220	13	2.69	0.429

d	Non-Inverted			Inverted			Unique		
SD-2-AFC	N	Mean	Standard Error	N	Mean	Standard Error	N	Mean	Standard Error
Soup	15	2.80	0.312	10	2.10	0.433	7	2.57	0.685
Kool-Aid	28	3.21	0.202	10	1.60	0.427	2	3.50	0.500
Tea	20	3.85	0.109	5	2.60	0.748	8	3.50	0.327
Tang	28	3.39	0.220	8	2.50	0.535	5	3.00	0.775

In the 2-AFC and Triangle tests with preview samples, subjects were divided into 'confident' and 'not confident' groups. Subjects who indicated that they could detect the difference between samples were categorized as confident. Those who indicated they could not detect the difference were categorized as not confident. With the Kool-Aid stimuli, the Triangle w/ Preview confident group performed significantly better ($p=0.0321$) than the not confident group. None of the other conditions for the Triangle w/ Preview or 2-AFC demonstrated a significant difference between confident and not confident groups. Additionally, there were no consistent non-significant trends (i.e. in some cases the not confident group performed slightly better than the confident group and in some cases the not confident group performed slightly worse than the confident group) (Table 9). The correlation between the proportion of subjects 'not confident' and the d' for the group was not as strong as with inversion ($R^2 = 0.403$). This suggests that subject confidence is not determined only by the difficulty of the difference they are looking for.

4.3 SAMPLE AND REPLICATE ORDER EFFECTS

Data were analyzed to look for sample order effects (i.e. WWS vs. SSW – W=Weaker Sample, S=Stronger Sample). For the two sample tests, t-tests were used to compare the two possible orders (WS and SW). For the three sample tests, ANOVA's were run comparing all six orders (SSW, SWS, WSS, WWS, WSW, and SWW). Where a significant effect was found in the ANOVA, a Tukey test was run. Analyses were conducted on the individual conditions (a total of sixteen comparisons) as well as over procedures (a total of eight comparisons). Significant effects were only found within the Tang conditions. Table 12 summarizes these effects. Where significant effects were found, they matched the effects predicted by Sequential Sensitivity Analysis (Rousseau and O'Mahony, 1996).

Table 12. Summary of significant order effects. All other comparisons were not significantly different at the 0.05 or 0.10 α-levels. *Significantly different at the 0.05 α-level by Tukey test.

Tang Condition	P-Value	Most Sensitive Order	Least Sensitive Order
All Triangles	0.0333	WWS*	WSS*
All 2-AFC's	0.0404	WS	SW
2-AFC w/ Preview	0.0218	WS	SW

Data were analyzed to determine if there was a significant difference between the results of four replicates (i.e. First replicate vs. Fourth replicate). ANOVA's were conducted comparing the four replicates for each condition (a total of sixteen ANOVA's). No significant effects were found at the 0.05 α-level. At the 0.10 α-level, only one significant difference was found. Within the Soup – Triangle condition, the first replicate was significantly worse than the second replicate at the 0.10 α-level.

CHAPTER 5

DISCUSSION

The main objective of this project was to determine the suitability of the SD-2-AFC as a replacement for the Triangle in complex food situations. While the SD-2-AFC was able to discriminate all of the products tested at a low alpha level ($p < 0.02$), so was the Triangle. Based only on this measure, there is no clear benefit of one test over the other. The SD-2-AFC requires an additional step, while requiring fewer samples than the Triangle (if replicates of both are conducted). Given the long track record of the Triangle, it is unlikely that the SD-2-AFC (which performs the same at best) would be a suitable replacement. However, if an experimenter was motivated to find a replacement for the Triangle, the SD-2-AFC is capable of detecting differences. The addition of a preview sample did not have a consistently positive or negative impact on performance. The rest of the analysis will focus on the potential reasons for the SD-2-AFC underperforming.

The underperformance of the atypical groups (those using attributes not in agreement with the plurality of subjects) and inverted groups (those using attributes opposite to those used by the plurality of subjects) compared to the typical/not inverted group may account for the SD-2-AFC's underperformance overall. This explanation is consistent with signal detection theory. When subjects are asked to choose their attribute, they taste the stronger and weaker sample. From this, subjects first decide what dimension to compare the samples on and then which sample is highest in that dimension. If a subject chooses a dimension that does not describe the difference they detect, they will be expected to perform at chance probability. For subjects who choose a dimension that does describe the difference they detect, there are three possibilities. A subject could pick the correct dimension but not be sensitive enough to discriminate the two samples. They will be expected on average to perform

at chance probability. A subject that can discriminate the samples could pick the correct dimension and correctly discriminate which sample is stronger in that dimension. They will be expected to perform above chance according to their individual sensitivity. The third possibility is for a subject that can discriminate the samples to pick the correct dimension but incorrectly discriminate which sample is stronger. This is expected to occur some proportion of the time. Even though subjects may reliably respond correctly above chance probability, according to SDT, it is expected that they respond incorrectly some of the time. This is due to "modal discriminal process for [a] given stimulus," wherein a subject's perception of a given stimulus is expected to vary over repeated exposures (Thurstone, 1927). This expected variation produces situations were a subject who can reliably discriminate two stimuli produces an incorrect response (Thurstone, 1927) (Figure 2). As d' increases, the likelihood of these incorrect responses is reduced (Frijters, 1979) (Figure 1). It is, however, never eliminated completely.

Figure 7 examines the scenario where an incorrect response occurs during attribute selection in the SD-2-AFC. The first and most likely scenario (on the left side) sees the subject correctly choosing the attribute and going on to perform three out of four tests correct. The second and less likely scenario (on the right side) sees the same subject with the same δ and same perceptions of the test samples choosing the inverted attribute. Because of this, every time where the non-inverted subject gets an answer correct, the inverted subject gets it incorrect. The error arises, not because of a lack of sensitivity or understanding on the part of the subject, but due to the variance in momentary perceptions of samples predicted by SDT. Those with high individual sensitivities will be very unlikely to be inverted. Those with very low individual sensitivities will have close to a 50% chance of being inverted (both samples appear identical so the subject must guess). It would also be expected that, as

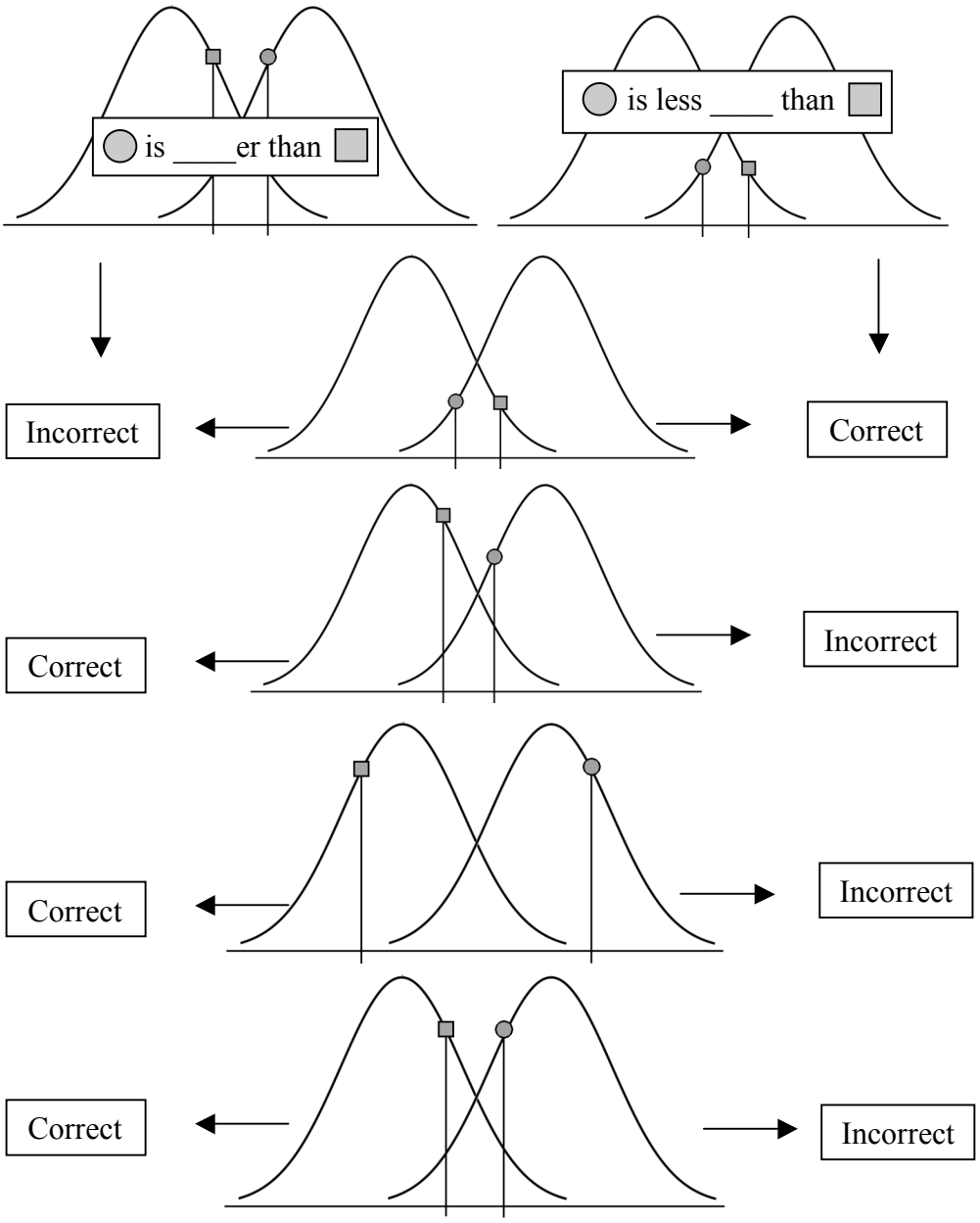


Figure 7. Hypothetical scenario where a subject chooses the correct criterion (left side) or the inverted criterion (right side) on the SD-2-AFC. The circle represents the momentary perception of the sample with more of the specific ingredient (e.g. sugar or citric acid). The square represents the momentary perception of the sample with less of the same ingredient. Note how the subject's proportion correct is different depending on the criterion even though the subject has identical perceptions of the samples during the testing replicates (bottom four curve sets).

the magnitude of the difference between stimuli increases, the proportion of subject inverted would decrease. The negative correlation ($R^2 = 0.937$) between the proportion of subjects inverted and the d' for the group (Figure 8) supports such a trend.

Inversion has not been mentioned it previous studies using methods that require subject-defined attributes. This is likely due to the key difference between the SD-2-AFC and the "warmed-up paired comparison." The SD-2-AFC intentionally does not have a full warm up to avoid increasing the sensitivity of the panel. However, in the "warmed-up paired comparison" subjects define the attributes after a warm up. This requires many repeated judgments of the difference between samples before the attribute is chosen. This reduces or eliminates altogether the chances of a subject being inverted. Subjects who would still be inverted after a full warm up are likely too insensitive to detect the difference anyway and would perform at chance probability. This suggests that increasing the number of preview samples would reduce the proportion of inverted subjects. However, every added preview sample would increase any warm up effect that are taking place. While one preview sample did not significantly increase the sensitivity of the subjects, it is unclear how many could be added without an unacceptable increase in alpha risk.

In the analysis of order effects, it is interesting to note that only one food product (Tang) elicited any significant effects. This consistency first suggests that the effect seen is genuine (as opposed to the Type I error expected over twenty four comparisons). Additionally, this suggests that some food products are more susceptible to sample order effects than others. The Tang stimuli fell in the middle as far as complexity and perceived difference between the weak and strong samples, suggesting these were not the most important factors in causing order effects. The same directions as to tasting order were given for all testing procedures, so subject

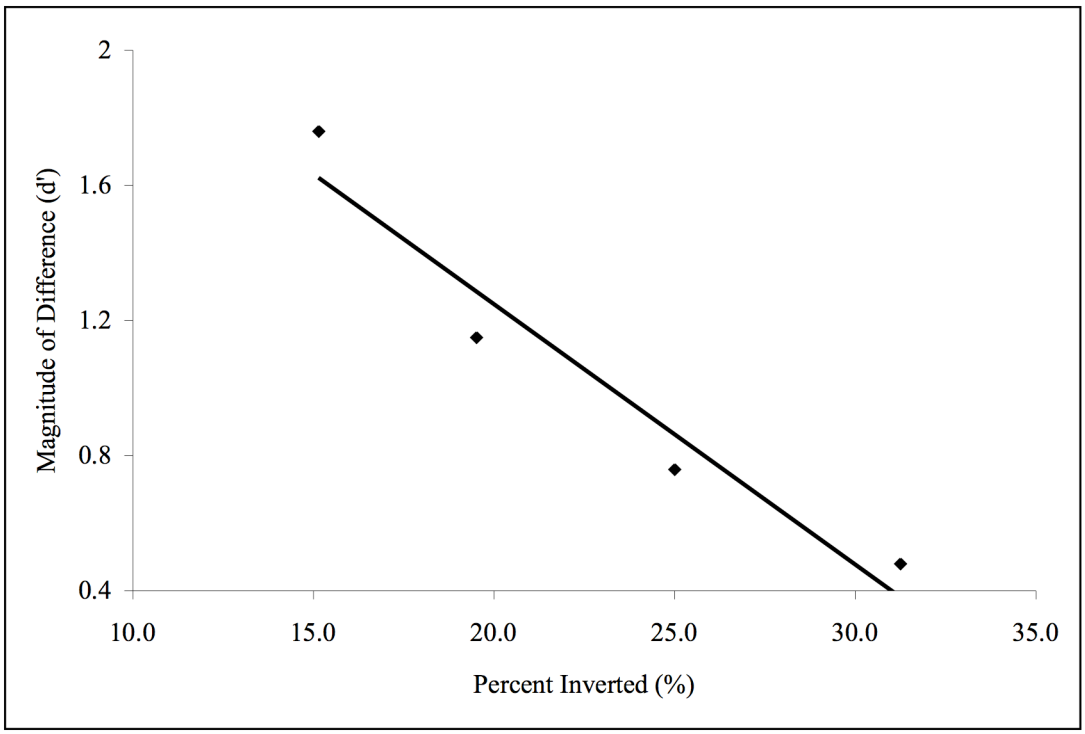


Figure 8. Correlation between d' and proportion of subjects inverted in the SD-2-AFC ($R^2 = 0.937$). Points are the group average d' for the SD-2-AFC for each food product.

compliance (i.e. tasting in the specified order) is not a likely cause of the effect. The direction of the significant results supported those predicted by studies involving SSA (Rousseau and O'Mahony, 1997; Rousseau et al., 1998; Lee and O'Mahony, 2007).

The replicate order effects did not show any significant trends. Due to this, and the large number of statistical comparisons (16 ANOVA's) performed, the significant effect detected is likely due to Type I error rather than a true effect. This is important because it justifies the pooling of all four replicates into one data point for each subject. If there was a significant directional change over replicates (i.e. subjects become significantly fatigued and perform worse on the last replicate than the first replicate) pooling the replicates would not be justified. Pooling the replicates allowed for smaller standard deviations without the added assumptions of the beta-binomial (as discussed in section 1.5.2). Dacremont and Sauvageot, 1997, found an increase in performance over replicates. In this study, however, the increase in performance was seen over eight replicates and was slight.

CHAPTER 6

CONCLUSIONS

The most important conclusion is that the SD-2-AFC does not appear to be a viable replacement for the Triangle test, nor does it appear to be a substitute for the 2-AFC. The reason for the underperformance of the SD-2-AFC appears to be inversion of the attributes by a portion of subjects. The most obvious solution to inversion would be including multiple preview samples before naming having subjects select their attribute. Unfortunately, this would revert the SD-2-AFC to the Warmed-Up Paired Comparison. While the Warmed-Up Paired Comparison is a more powerful procedure, its use where an untrained panel representative of a group of consumers is desired would increase the likelihood of Type I error (O'Mahony et al., 1988). The purpose of developing the SD-2-AFC was to provide a new procedure for use with an untrained panel representative of a group of consumers. The warm-up effects seen in the Warmed-Up Paired Comparison result in a panel more sensitive than the group of consumers they are meant to represent. It is interesting to note that the presence of inversion is consistent with SDT. This study, therefore, provides more support for the use of SDT as a theoretical model for discrimination testing.

APPENDIX I.
INFORMED CONSENT FORM

Revised consent form 9/10/07
INFORMED CONSENT

This is a study to investigate the sense of taste. You will be tasting products commercially available and intended for human consumption.

The experiment will consist of 1 session. And will take approximately 10 minutes. After the study is completed, one $100 and two $50 ~~Best Buy gift cards~~ cash amounts will be raffled off to all participants in the study.

You will not be paid for participating in the experiment. If you are a student, no class credit is involved.

Your participation is strictly voluntary. You have the right to leave the experiment at any time you wish, without any penalty or hard feelings. Such a decision will not influence any other relationship that you may have to Cornell University, the experimenter (Dr. Lawless), his staff or students in any way.

There are no right or wrong answers in these tests. It is your perceptions about the stimulus materials that we are interested in. After the experiment, your data will be kept in a locked room. In any electronic records, you will be identified only by a code number. Your personal data will never be displayed in any presentation or publication with your identity revealed by name or initials. If you are recruited over the internet, it is possible that emails can be intercepted. Personal data will not be requested over the internet. Please ask any questions you have about the study at this time. If you have questions at any later time please contact Harry Lawless, 255 - 7363, or Scott McClure, stm27@cornell.edu.

By signing below, I indicate that I am participating in this study voluntarily. I understand that I have the right to withdraw from the experiment at any time, without penalty. I also indicate that to the best of my knowledge, I have a normal sense of taste and smell, and that I have none of the following conditions: any chronic health problem, respiratory disease such as a cold or asthma, respiratory allergies such as hayfever, and/or food allergies. All my questions about the experiment have been answered to my satisfaction. I am between the ages of 18 and 65, inclusive.

Contact Information: Harry Lawless, 255 - 7363, htl1@cornell.edu,
Scott McClure, stm27@cornell.edu.

You may contact the Institutional Review Board for Human Participants (IRB)
www.irb.cornell.edu / irbhp@cornell.edu / 607-255-5138

This consent form will be kept by the researcher for at least three years beyond the end of the study and was exempted by ORIA on 9/10/2007.

Name
_____Date_____

Signature_____

APPENDIX II. EXAMPLE TEST BALLOTS

A. Example of Triangle Test Ballot

Subject 27.1

Fruit Drink Taste Test

Two of these samples are the same and one is different. Taste the samples from left to right and circle the number of the **different sample.**

Circle one sample in each box.
Rinse between boxes.

801	**412**	**387**
610	**248**	**369**
159	**904**	**198**
745	**977**	**134**

B. Example of 2-AFC Test Ballot

Subject 27.5

Fruit Drink Taste Test

These two samples are different. Taste the samples from left to right and circle the number of the **sweeter sample.**

Circle one sample in each box.
Rinse between boxes.

291	**178**
450	**209**
248	**369**
486	**593**

C. Example of SD-2-AFC Test Ballot

Subject 27.7

Fruit Drink Taste Test

These two samples are different. Taste the samples from left to right and circle the number

of the _____ **sample.**

Circle one sample in each box.
Rinse between boxes.

291	**178**
450	**209**
242	**509**
486	**593**

APPENDIX III. EXAMPLE PREVIEW BALLOTS

A. Example of 2-AFC Preview Ballot

Subject __49.6__

Sample 964 is sourer than Sample 578.
Taste both samples circle a statement below

Circle One

I can tell that 964 is sourer than 578

I cannot tell that 964 is sourer than 578

B. Example of Triangle Preview Ballot

Subject __53.2___

Sample 964 is different than Sample 578.
Taste both samples circle a statement below

Circle One

I can tell that 964 is different than 578

I cannot tell that 964 is different than 578

C. Example of SD-2-AFC Preview Ballot

Subject ___53.7___

Sample 964 is different than Sample 578.
Taste both samples and fill in the blank below describing the difference

Sample 964 is

_____

than Sample 578

APPENDIX IV.
ESTIMATION OF VARIANCE OF d' USING THE METHOD DETERMINED BY BI ET AL. (1997)

Below are the calculations required to determine if any of the procedures used to test the Tea Samples produced significantly different d's.

	N	Mean Correct	Proportion Correct	d'	B Value	Variance (S^2)
Triangle	34	3.12	0.78	2.98	7.670	0.2256
Triangle w/ Preview	34	3.44	0.86	3.61	10.183	0.2995
SD-2-AFC	33	3.58	0.895	1.76	5.6344	0.1707

1. Mean Correct determined from raw data
2. Mean Correct converted to Proportion Correct
   a. $\Pr opotion\, correct = \frac{Mean\, correct}{Number\, of\ tests}$
3. d' determined using proportion correct and the tables provided by Ennis (1993)
4. B Value determined using d' and the tables provided by (Bi et al., 1997)
5. Variance of d' calculated using B and N
   a. $Variance\, of\ d' = \frac{B}{N}$
6. Using the estimates of variance and d', X^2 can be calculated
   a. $$\exp ected\, d'\, value = d'_{\exp} = \frac{\frac{d'_{Triangle}}{S^2_{Triangle}} + \frac{d'_{Triangle\, w/\Pr eview}}{S^2_{Triangle\, w/\Pr eview}} + \frac{d'_{SD-2-AFC}}{S^2_{SD-2-AFC}}}{\frac{1}{S^2_{Triangle}} + \frac{1}{S^2_{Triangle\, w/\Pr eview}} + \frac{1}{S^2_{SD-2-AFC}}}$$
   b. $$\exp ected\, d'\, value = d'_{\exp} = \frac{\frac{2.98}{0.2256} + \frac{3.61}{2.995} + \frac{1.76}{0.1707}}{\frac{1}{0.2256} + \frac{1}{2.995} + \frac{1}{0.1707}} = 2.6101$$
   c. $$x^2 = \frac{(d'_{Triangle} - d'_{\exp})^2}{S^2_{Triangle}} + \frac{(d'_{Triangle\, w/\Pr eview} - d'_{\exp})^2}{S^2_{Triangle\, w/\Pr eview}} + \frac{(d'_{SD-2-AFC} - d'_{\exp})^2}{S^2_{SD-2-AFC}}$$

d. $$\boxed{x^2 = \frac{(2.98-2.6101)^2}{0.2256} + \frac{(3.61-2.6101)^2}{2.995} + \frac{(1.76-2.6101)^2}{0.1707} = 8.1774}$$

e. $$\boxed{p(x^2 = 8.1774, d.f. = 2) = 0.0167}$$

7. Since a difference was found, Z-values for each comparison must be determined

8. $$Z_{Triangle-Triangle\,w/\,preview} = \frac{\left|d'_{triangle} - d'_{triangle\,w/\,preview}\right|}{\sqrt{S^2_{triangle} + S^2_{triangle\,w/\,preview}}} = \frac{|2.98-3.61|}{\sqrt{0.2256+0.2995}} = 0.8694$$

$$Z_{SD-2-AFC-triangle\,w/\,preview} = \frac{\left|d'_{SD-2-AFC} - d'_{triangle\,w/\,preview}\right|}{\sqrt{S^2_{SD-2-AFC} + S^2_{triangle\,w/\,preview}}} = \frac{|1.76-3.61|}{\sqrt{0.1707+0.2995}} = 2.698$$

$$Z_{triangle-SD-2-AFC} = \frac{\left|d'_{triangle} - d'_{SD-2-AFC}\right|}{\sqrt{S^2_{triangle} + S^2_{SD-2-AFC}}} = \frac{|2.98-1.76|}{\sqrt{0.2256+0.1707}} = 1.938$$

9. $p(Z_{Triangle-Triangle\,w/\,preview}) = 0.1923$
   $p(Z_{SD-2-AFC-Triangle\,w/\,preview}) = 0.003488$
   $p(Z_{Triangle-SD-2-AFC}) = 0.02631$

10. From this it can be shown that the SD-2-AFC performed significantly worse than the Triangle and Triangle w/ preview at the 0.05 alpha level.

REFERENCES

Amoore, J.E. 1977. Specific Anosmia and the concept of primary odors. Chemical Senses and Flavor, 2, 267-281.

Amoore, J.E.; Venstrom, D.; Davis, A.R. 1968. Measurement of Specific Anosmia. Perceptual and Motor Skills, 26, 143-164.

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

Aust, L.B.; Gacula, M.C. Jr.; Beard, S.A.; Washam, R.W. II. 1985. Degree of Difference Test Method in Sensory Evaluation of Heterogeneous Product Types. Journal of Food Science, 50, 511-513.

Bi, J. 2006a. Statistical Analyses for R-Index. Journal of Sensory Studies, 21, 584-600.

Bi, J. 2006b. Sensory Discrimination Tests and Measurements: Statistical Principles, Procedures and Tables. Blackwell Publishing, Oxford. 1-20.

Bi, J. and Ennis, D.M. 1999a. Beta-Binomial tables for replicated difference and preference tests. Journal of Sensory Studies, 14, 347-368.

Bi, J. and Ennis, D.M. 1999b. The power of sensory discrimination methods used in replicated difference and preference tests. Journal of Sensory Studies, 14, 289-302.

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

Bradley, R.A. 1963. Some Relationships among sensory difference tests. Biometrics, September, 385-397.

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

Brockhoff, P.B. and Schlich, P. 1998. Handling replications in discrimination tests. Food Quality and Preference, 9, 303-312.

Byer, A.J. and Abrams, D. 1953. A comparison of the triangular and two-sample taste-test methods. Food Technology, 7, 185-187.

Dacremont, C. and Sauvageot, F. 1997. Are replicate evaluations of triangle tests during a session good practice? Food Quality and Preference, 8, 367-372.

Dacremont, C.; Sauvageot, F.; Duyen, T.HA. 2000. Effect of assessors' expertise level on efficiency of warm-up for triangle tests. Journal of Sensory Studies, 15, 151-162

David, H.A. and Trivedi, M.C. July 1962. Research on Order Statistics and the Design of Experiments: Pair, Triangle, and Duo-Trio Tests. Virginia Polytechnic Institute Department of Statistics Technical Report Number 55. Blacksburg, VA.

Delwiche, J. and O'Mahony, M. 1996. Flavour discrimination: and extension of Thurstonian 'paradoxes' to the tetrad method. Food Quality and Preference, 7, 1-5.

Delwiche, J.F.; Buletic, Z.; Breslin, P.A.S., 2001. Covariation in individuals' sensitivies to bitter compounds: Evidence supporting multiple receptor/transduction mechanisms. Perception and Psychophysics, 63, 761-776.

Dessirier, J.M and O'Mahony, M. 1999. Comparison of d' values for the 2-AFC (paired comparison) and 3-AFC discrimination methods: Thurstonian models, sequential sensitivity analysis and power. Food Quality and Preference, 10, 51-58.

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

Ennis, D.M. and Bi, J. 1998. The Beta-Binomial model: accounting for inter-trial variation in replicated difference and preference tests. Journal of Sensory Studies, 13, 389-412.

Ennis, D.M. and Mullen, K. 1985. The effect of dimensionality on results from the triangular method. Chemical Senses, 10, 605-608.

Ennis, D.M.; Palen, J.J.; Mullen, K. 1988. A multidimensional stochastic theory of similarity. Journal of Mathematical Psychology, 32, 449-465.

Ferdinandus, A.; Oosterom-Kleijngeld, I.; Runneboom, A.J.M. 1970. Taste Testing. MBAA Technical Quarterly, 7, 210-227.

Filipello, F. 1956. A critical comparison of the two-sample and triangular binomial designs. Journal of Food Science, 21, 235–241

Frijters, J.E.R. 1979. The paradox of discriminatory nondiscriminators resolved. Chemical Senses and Flavour, 4, 355-358.

Frijters, J.E.R. 1982. Expanded tables for conversion of a proportion of correct responses (Pc) to the measure of sensory difference (d') for the triangular method and the 3-alternative forced-choice procedure. Journal of Food Science, 47, 139-143.

Frijters, J.E.R. 1988. Sensory difference testing and the measurement of sensory discriminability. In Sensory Analysis of Foods Second Edition, ed Piggott, J.R. Elsevier Applied Science, New York.

Giboreau, A.; Dacremont, C.; Egoroff, C.; Guerrand, S.; Urdapilleta, I.; Candel, D.; Dubois, D. 2007. Defining sensory descriptors: Towards writing guidelines based on terminology. Food Quality and Preference, 18, 265-274.

Green, D.M and Swets, J.A. 1966. Signal Detection Theory and Psychophysics. John Wiley and Sons, Inc, New York.

Gridgeman, N.T. 1959. Pair Comparison, with and without ties. Biometrics, 15, 382-388.

Hacker, M.J. and Ratcliff, R. 1979. A revised table of d' for M-alternative forced choice. Perception and Psychophysics, 26, 168-170.

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

Hautus, M.J. and Irwin, R.J. 1995. Two models for estimating the discriminability of foods and beverages. Journal of Sensory studies, 10, 203-215.

Helm, E. and Trolle, B. 1946. Selection of a Taste Panel. Wallerstein Laboratories Communications, 9, 181-194.

Hopkins, J.W. and Gridgeman, N.T. 1955. Comparative Sensitivity of Pair and Triad Flavor intensity difference tests. Biometrics, March, 63-68.

Jiamyangyuen, S.; Delwiche, J.F.; Harper, W.J. 2002. The impact of wood ice cream sticks' origin on the aroma of exposed ice cream mixes. Journal of Dairy Science, 85, 355-359.

Kim, H.J.; Jeon, S.Y.; Kim, K.O.; O'Mahony, M. 2006. Thurstonian models and variance I: experimental confirmation of cognitive strategies for difference tests and effects of perceptual variance. Journal of Sensory Studies, 21, 465-484.

Kunert, J. and Meyners, M. 1999. On the triangle test with replications. Food Quality and Preference, 10, 477-482.

Lawless, H.T. and Heymann, H. 1998. Sensory Evaluation of Food: Principles and Practices. Springer Science, NY.
A) Chapter 1: Introduction and Overview, 1-27.
B) Chapter 3: Principles of Good Practice. 83-115.
C) Chapter 4: Discrimination Testing 116-139.
D) Chapter 5: Discrimination Theories and Advanced Topics 140-172.
E) Appendix I: Basic Statistical Concepts for Sensory Evaluation, 647-678.
F) Appendix V: Statistical Power and Test Sensitivity 754-782.

Liggett, R.E. and Delwiche, J.F. 2005. The Beta-Binomial model: variability in Overdispersion across methods and over time. Journal of Sensory Studies, 20, 48-61.

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

MacRae, A.W. 1995b. Visualizing the difference between triangle and 3AFC judgments. Food Quality and Preference, 6, 315-320.

Mata-Garcia, M.; Angulo, O.; O'Mahony, M. 2007. On Warm-Up. Journal of Sensory Studies, 22, 187-193.

O'Mahony, M. 1995. Who told you the triangle test was simple? Food Quality and Preference, 6, 227-238.

O'Mahony, M.; Thieme, U.; 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.

O'Mahony, M.A.P.D. 1979. Short-cut signal detection measures for sensory analysis. Journal of Food Science, 44, 302-303.

O'Mahony, M. and Goldstein, L.R. 1986. Effectiveness of sensory difference tests: sequential sensitivity analysis for liquid food stimuli. Journal of food science, 51, 1550-1553.

O'Mahony, M. and Rousseau, B. 2002. Discrimination testing: a few ideas, old and new. Food Quality and Preference, 14, 157-164.

O'Mahony, M.; Goldenberg, M.; Stedmon, J.; Alford, J. 1979. Confusion in the use of the taste adjectives 'sour ' and 'bitter'. Chemical Senses, 4, 301-318.

O'Mahony, M.; Masuoka, S.; Ishii, R. 1994. A theoretical note on difference tests: models, paradoxes, and cognitive strategies. Journal of Sensory Studies, 9, 247-272.

O'Mahony, M.; Wong, S.Y.; Odbert, N. 1985. Sensory evaluation of navel oranges treated with low doses of gamma-radiation. Journal of Food Science, 50, 639-646.

O'Mahony, M.; Wong, S.Y.; Odbert, N. 1986. Sensory difference tests: Some rethinking concerning the general rule that more sensitive tests use fewer stimuli. LEBENSMITTEL-WISSENSCHAFT UND-TECHNOLOGIE-FOOD SCIENCE AND TECHNOLOGY, 19, 93-94.

O'Mahony, M.A.P.D.E. and Odbert, N. 1985. A comparison of sensory difference testing procedures: sequential sensitivity analysis and aspects of taste adaptation. Journal of food science, 50, 1055-1058.

Pelletier, C.A.; Lawless, H.T.; Horne, J. 2004. Sweet–sour mixture suppression in older and young adults. Food Quality and Preference, 15, 105-116.

Peryam, D.R. and Swartz, V.W. 1950. Measurement of Sensory differences. Food Technology, 10, 390-395.

Pfaffmann, C. 1953. Variables Affecting Difference Tests. 1953 Quartermaster Corp symposium.

Priso, H.E.; Danzart, M.; Hossenlopp, J. 1994. A statistical analysis of difference tests with replications. Journal of Sensory Studies, 9, 121-130.

Radkins, A.P. 1957. Some statistical considerations in organoleptic research: triangle, paired, duo-trio tests. Food Research, 22, 259-265.

Rainey, B.A. 1986. Importance of reference standards in training panelists. Journal of Sensory Studies, 1, 149-154.

Rousseau, B. 2001. The Beta-Strategy: An alternative and powerful cognitive strategy when performing sensory discrimination tests. Journal of Sensory Studies, 16, 301-318.

Rousseau, B. 2006. Indices of Sensory Difference: R-Index and d'. The Institute for Perception IFPress, 9, 2-3.

Rousseau, B. and Ennis, D. 2007. Why Proportion of discriminators is method-specific. The Institute for Perception IFPress, 10, 2-3.

Rousseau, B. and O'Mahony, M. 1997. Sensory difference tests: Thurstonian and SSA predictions for vanilla flavored yogurts. Journal of Sensory Studies, 12, 127–146.

Rousseau, B.; and Ennis, D.M. 2001. A Thurstonian model for the dual pair (4IAX) discrimination method. Perception and Psychophysics, 63, 1083-1090.

Rousseau, B.; Rogeaux, M.; O'Mahony, M. 1999. Mustard discrimination by same-different and triangle tests: aspects of irritation, memory and t criteria. Food Quality and Preference, 10, 173-184.

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.

Tarone, R.E. 1979. Testing the Goodness of Fit of the Binomial Distribution. Biometrika, 66, 585-590.

Theime, 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.

Thurstone, L.L. 1927. A law of comparative judgment. Psychological Review, 34, 273-286.

Ura, S. 1960. Pair, Triangle and Duo-Trio Test. Reports of Statistical Application Research. Union of Japanese Scientists and Engineers, Tokyo, 7, 1-13.