Appendix A: Overview of Sensory Evaluation

> *Successful sensory testing is driven by setting clear objectives, developing robust experimental strategy, applying appropriate statistical techniques, adhering to good ethical practice and successfully delivering actionable insights that are used to inform decision-making.*
>
> Kemp et al. (2009)

A.1 Introduction

Sensory evaluation is a child of the industries that manufacture beverages, foods, and consumer products. The techniques have evolved since the mid 1990s, and many of them still are practiced in the forms in which they were first published. The goal of these kinds of tests was to get an insight into human perception of the products that could be used to guide management decisions. Often, these decisions concerned the development and introduction of a new processed food to the consumer market. The tests were also designed for purposes of quality control and stability (shelf-life) testing. The main idea was to provide information that would lower the risk in making decisions about a product, such as whether to sell it. The specific questions involved were: (1) Was it was the same as or different than an existing product? (2) What were the perceived characteristics of this product? (3) Would consumers like it? To answer these questions, the methods of discrimination testing, descriptive

Quantitative Sensory Analysis: Psychophysics, Models and Intelligent Design, First Edition. Harry T. Lawless.

analysis, and affective testing were developed. They will be discussed briefly in the sections that follow. More detailed information can be found in textbooks on the subject, including Lawless and Heymann (2010), Stone et al. (2004), Meilgaard et al. (2006), and the shorter works by Chambers and Wolf (1996) and Kemp et al. (2009).

I have assumed that most readers of this work will be familiar with the basic types of sensory tests, how they are conducted, and how the data are analyzed. However, not everyone who picks up this book will have previous experience or formal studies of the subject matter, and so a short introductory summary is given in this appendix in order to provide a background for such individuals. It is also assumed that the reader is familiar with basic statistical terminology, such as the null hypothesis, mean, standard deviation, and correlation. If not, the appendices to Lawless and Heymann (2010) give a useful introduction focused on statistical applications for sensory evaluation.

A.1.1 The Central Dogma

The central dogma of sensory evaluation is that the test method must be designed to match the objectives of the test. If one wishes to know whether or not there is a perceivable difference between two products, then a discrimination or simple difference test is needed. If consumer acceptability or preference is unknown, then a consumer test is required. The metaphor is often used that the methods are part of a "sensory toolbox." You would not use a hammer when a screwdriver is needed. So the tool is fit to the problem at hand.

As a corollary to this central concept, certain types of assessors must be used in each type of test. For a descriptive analysis panel, persons must have been screened to make sure they have normal sensory acuity, proper motivation, and no health issues that would contraindicate their consumption of the products. However, they do not need to be regular purchasers of this product in question. They are merely functioning as analytical instruments; sensory meters of sorts. The opposite is true for consumer tests. They must be regular and perhaps frequent users of the product or the product category being tested. However, they are more or less randomly sampled from the population of such users, and hence their sensory acuity is not a criterion for participation. A third principle is that we do not ask the wrong questions of the wrong panel. If we have a trained analytical descriptive panel, we do not ask them whether they like the product. That is not their job. Conversely, we are very careful about asking consumers for analytical diagnostic evaluations of specific sensory properties. They have not been trained to use a common vocabulary, and so asking questions about an attribute like astringency (which often must be taught to a panel with examples) would not make sense.

A.1.2 Blind Testing and Independent Judgments

Two guiding principles are as important as the central dogma. The first is blind testing. That is, the tester must not be aware of the identity of the products in the test, nor which one is a new test product versus a control sample. They are given only enough information to make a judgment in the proper frame of reference. We do not tell them the product concept. An example would be this: "We have a new, whole wheat, high fiber, low fat microwavable frozen pizza that your kids can safely prepare for themselves after school." That kind of concept testing is the province of marketing research. Instead, the product comes to the taste testing booth with the designation "Pizza # 357." Note that a random three-digit blinding code is used to label the sample. The second guiding principle is that of independent judgments. We do not want the assessors to discuss the properties and come to a group decision.

It is their individual opinions and data that matter. The statistical assumption is that the observations are independent. So their judgments must not be influenced by other persons in the test.

A.1.3 Facilities and Controls

In order to conform to the principles of blind testing and independent judgments, sensory testing in industry has followed some general practices. Independent judgment is facilitated by having assessors separated. This is usually achieved by seating them in a private booth, with barriers between booths to prevent interaction with other persons. Sensory evaluation booths are often connected via a pass-through window or other opening to a test kitchen or preparation area. The opening usually has a door that remains closed most of the time, so that the tester cannot see the prep area or get an unwanted hint about the identity of the products. The identity of products is further obscured by labeling the samples with random three-digit codes.

Samples are prepared in a uniform manner, and should have the same volume (size) and serving temperature. They should only differ in the variable under investigation. In other words, there should be no unwanted systematic or random variation that would add error variability to the situation or cause perception of a spurious or irrelevant difference. Of course, many products such as a snack chips have normal variation and this is to be expected. But it would be very bad practice to have the test product of a different size than the control product, unless size was in fact the variable under study. A laboratory manual with standard procedures should be part of any ongoing testing program, so that the sample preparation is standardized and repeatable.

A.2 Discrimination and Simple Difference Tests

A.2.1 Objectives

The primary goal of a discrimination test is to provide scientific evidence that two products are perceptually different. This can be rephrased as "there are at least some people in the population that can tell the difference." An important part of the testing is that it must be objective. Differentiation is not a matter of opinion, but a behavioral demonstration of the ability to discriminate is required. For this reason, most of these tests take the form of a multiple choice test, with a known chance probability level of getting the correct answer by guessing. Performance above these levels is considered evidence for a difference. When the level of performance is high enough, and the number of judges is sufficient, then a statistical test can be performed to show that the observed proportion of correct answers would only occur 5% of the time or less, under a true null hypothesis. In the case of a multiple choice format, the null states that the proportion correct in the general population equals the chance probability level. Note that this is a mathematical equality, and not a verbal statement like "there is no difference."

Discrimination tests can also be used to amass evidence that two products are equal or have some acceptable degree of sensory similarity. This is a trickier decision, and is more fully discussed in Chapter 6 on equivalence testing. A simple failure to find any statistically significant difference is weak evidence and often ambiguous. A failure to reject the null can happen for lots of reasons, including (1) you did a sloppy test that introduced sources of unwanted variability, (2) you did not test enough judges to provide a powerful and sensitive test, and (3) there really is no difference.

A.2.2 Participants

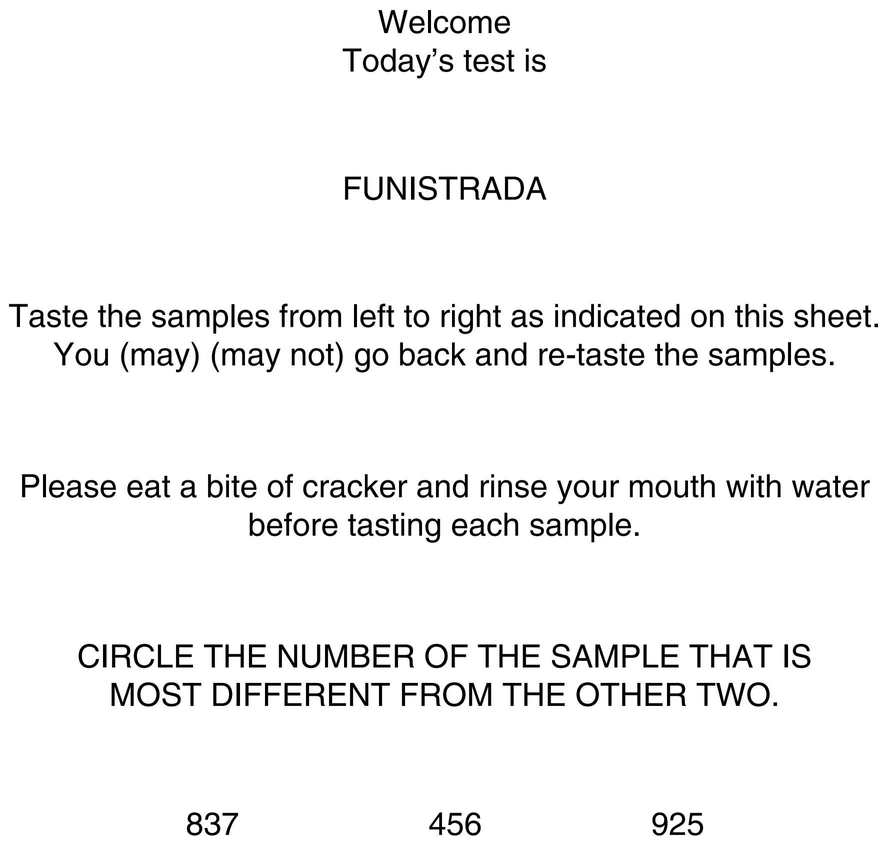
The most common scenario for a discrimination testing panel in a major food company is a reservoir of employees, often in a research setting. A subset of the employee panel can be called for any given test, until the necessary quota is reached. The panel size ranges from about 40 to 80, with the larger panels being used for equivalency testing. The panelists need not be users of the product, but they must have no objections to consuming the kind of food in that day's test. Often, they are screened for basic sensory acuity, and for any health reasons or dietary restrictions (religious, medical, or otherwise) that would prohibit them from eating that product.

In another scenario, consumers of that product type are recruited for the discrimination test. They are generally not screened for basic sensory function. Thus, this kind of panel is generally less sensitive than a screened employee panel, who may become quite discriminating due to the years of practice they may get taking taste tests. A more discriminating group of testers can provide better insurance against missing a difference (type II error) but can result in some differences being detected that most consumers might not see. A consumer panel, on the other hand, is more predictive of the discrimination abilities of the general population. The results have a greater chance of type II error, but a lower risk of finding a spurious difference; that is, a false alarm or type I error. The risks associated with each of these errors needs to be considered in setting up the type of panel the client needs. The employee panel is often thought of as a "safety net." That is, if a screened and discriminating panel cannot tell the difference between the products under controlled conditions, a consumer panel in the real world is unlikely to see the difference either. This logic seems good, but it is not airtight, because every test has some probability of error.

A.2.3 Methodologies

There are four main categories of methods used for simple difference tests. They are sorting, matching, forced-choice or attribute-specific, and response choice tests. The first three require the panelist to choose a specific product from a group of blind-coded items. The response choice tests involve choice of a response option, rather than pointing to a product. These have different statistical analyses and require a control product or control pair so as to provide a baseline of response to control for response bias. Each of these categories will be described next, and details of methods and analysis can be found in Lawless and Heymann (2010: chapter 4).

The most common sorting test is the triangle procedure, which has a long history in sensory testing. Three products are presented with random blinding codes, and two of them are duplicate items. The task is usually phrased as "choose the one item that is most different from the other two." A sample ballot for a triangle test is shown in Figure A.1. This task is also called an oddity task, for obvious reasons (choose the odd sample). The chance probability is one-third. Assuming there are two products in the test (e.g., a control product and a test item), the identity of the duplicate item is counterbalanced so that half the panelists get the control as a duplicate and half get the test product as the duplicate. Products will also be presented in different random orders to each person, or in



Welcome
Today's test is

FUNISTRADA

Taste the samples from left to right as indicated on this sheet.
You (may) (may not) go back and re-taste the samples.

Please eat a bite of cracker and rinse your mouth with water
before tasting each sample.

CIRCLE THE NUMBER OF THE SAMPLE THAT IS
MOST DIFFERENT FROM THE OTHER TWO.

837 456 925

Figure A.1 A typical ballot used in a triangle test.

different positions on a tray. A second kind of sorting test is the tetrad. The tetrad test has four items, with two pairs of duplicates. The job of the panelist is to sort them into two piles on the basis of similarity, so the duplicates are sorted together correctly. This also has a chance probability of one-third. These are good tests for overall difference, when the nature of the difference is unknown before the test, or is likely to be a complex set of changes across several attributes. Recent modeling suggests the tetrad test is generally more sensitive than the triangle (see Chapter 5), but it does not have the track record of decades of use, and obviously involves tasting one extra sample. In both of these tests a response is forced. That is, the panelist must respond by guessing if unsure. They are not allowed to say, "I can't tell" or "I have no response."

Matching tests are also quite popular for testing for overall difference. The duo–trio test is the oldest. In this procedure, a reference sample is presented to the tester, sometimes following a warm-up sample. After the person has a chance to inspect the reference item, two test items are presented, one of which matches the reference, and the panelist is to choose the item that is most similar to the reference. A second kind of test, the ABX test, is common in psychophysics and is the reverse of the duo–trio. Both test and control products are presented as references or whatever two versions of the product are sent to the sensory lab for testing. After the inspection period, one of the two items is presented with a blind code to each panelist, and he or she must match the third item to the correct reference. The identity of the third item, of course, is varied across panelists so the test product and control product appear an equal number of times. In the dual standard test, once again both items are presented as reference samples, and then both items are presented with blind codes, to be matched to the correct reference. In all three of these tests, choosing the correct match has a chance probability of one-half.

The third major category of difference tests is called forced choice. One of the test items is considered to be a priori higher in some attribute and the others are duplicated versions of the less intense or baseline products. The panelist is instructed to "choose the product that is highest in attribute X." If there is one test item and two control or baseline products, the task is called a three-alternative forced choice test or 3-AFC. If there is one test item and one baseline, it is called a 2-AFC or paired comparison test for differences. More baseline products can be added, resulting in 4-AFC and so on. This changes the chance probability level and increases the difficulty of the test. Because the sorting and matching tests also involve a forced choice, this terminology is somewhat unfortunate. For this reason, some sensory scientists prefer the term "attribute-specific" tests. The AFC tests are generally considered to be more sensitive than the nonspecific tests (see Chapter 4). However, the attribute that is changing must be known, and it must be a sensory property that the panelists understand and can easily recognize (such as sweet taste).

The fourth category involves a choice of response, rather than pointing to a product in a test set. These are the A–not-A test and the same–different test. Assume once again that there is a control product and a test product. If we designate the control product as "A" and the test product as "B," the requirements of these tests become clear. In the A–not-A test, the panelist is given a chance to inspect the control product, in order to learn what "A" is like. This inspection may be more or less stringent and time consuming, but the familiarity of the panelist with "A" must not be assumed. The panelist is then given blind-coded test products, one at a time, and must decide after each one whether it is an example of "A" or something else (hence not-A). Products A and B are given an equal number of times. The presence of both A and B is important to get a baseline of response frequencies for each item (you cannot just give B and see what people call it). If the two products are presented once to each panelist, the response matrix forms a 2 × 2 chart, and the McNemar test is an appropriate statistical test. If they are only presented once (half the panelists see A and half see B) a χ^2 test is appropriate, or a Z-test on proportions (see Bi (2006) for other models).

The same–different test requires a similar baseline condition. Now products are presented in pairs, and the response options are "same" or "different." The question, of course, is the frequency with which the AB pair is called "different," but this must be compared with the frequency with which the AA pair (control paired with itself) is called "different."

A similar requirement is present in any test with a rated degree of difference (also called a DOD test). Instead of two response options, the panelist may now signify some graded degree of perceived difference for each pair. This can be done on a line scale, or a category scale with verbal anchors ranging from "exact match" to "extremely different" or some similar wording. Now the question becomes whether the rated difference is higher for an AB pair than for an AA pair (or BB if that is a control condition). If each panelist sees each kind of pair once, a paired t-test is appropriate for scaled data. Similar designs and analyses are done for scales with sureness or certainty ratings, as discussed in Chapter 9.

A.2.4 Statistical Analysis

Because the chance probabilities and null hypotheses are known, the choice tests are straightforward to analyze. Under a true null, the expected data set would follow a binomial distribution. However, since the panel size is usually around 50 persons or above, the *Z*-score approximation to the binomial distribution can be used. The critical *Z*-value for a one-tailed test at $p=0.05$ is 1.645. So the difference between the observed proportion correct and the chance proportion must satisfy the following inequality:

$$\frac{\left(P_{\text{observed}} - P_{\text{chance}}\right) - \dfrac{1}{2N}}{\sqrt{p(1-p)/N}} \geq 1.645 \qquad \text{(A.1)}$$

where *N* is the number of panelists and *p* is the chance probability level. The term $1/2N$ is a correction for continuity. If P_{observed} is equal to the count of the number correct *X* divided by *N*, we get the following formula, which is preferred by some authors and students:

$$\frac{(X - Np) - 0.5}{\sqrt{Np(1-p)}} \geq 1.645 \qquad \text{(A.2)}$$

Because the chance probability level is known for each test, simple significance tables are found in most textbooks, showing the minimum number of correct choices that are required (*X*) as a function of *N* in order to satisfy the inequality. The tests are one tailed, because only performance above chance is predicted, and the alternative hypothesis states that the population proportion correct would be greater than the chance probability level (rather than not-equal-to). Analyses for replicated tests can use the beta binomial models, as discussed in Chapter 5 and in Bi (2006).

A.3 Descriptive Analysis

A.3.1 Objectives

Simple difference tests provide evidence that at least some persons can tell two products apart. But they do not tell you in what ways the products differ. This is the goal of a descriptive analysis. Descriptive analysis provides a complete specification of the sensory properties of a food or consumer product. This is achieved by two important processes: first, understanding the terms that will be used to describe the product and, second, specifying the strength or perceived intensity of each attribute. Note that this is a psychophysical model in most cases. The attribute varies from weak to strong sensations in that class of products, or perhaps from no sensation of that type to a strong sensation. Descriptive analysis is not concerned with liking or preference for the product, nor is it concerned with feelings or emotions. It answers the “what” in what is different about these products.

The tool is a versatile one. It can be used for any situation in which an ingredient, processing, or packaging variable has changed, as well as for shelf life and stability testing. Multiple products can be submitted for testing. Replicated measurements are common. Generally, a panel will be trained only to evaluate one class or category of products. So in a large food manufacturer, there may be several different panels operating for different product lines. Of course, the participants must be trained to understand the terms that are used to describe the products. This is often done through the use of examples or reference standards. In some versions of these techniques, they are also calibrated to use the intensity scale by means of reference examples of items that are weak or strong in that specific attribute. Statistical analysis is necessary and is discussed further below. Further information on descriptive analysis procedures can be found in Lawless and Heymann (2010: chapter 10).

A.3.2 Participants

Setting up a descriptive panel takes a lot of time for screening and training. Once the panel is in place, it is typically used several times a week. So, two important parts of qualification for panel work are (1) the motivation to participate and (2) supervisory approval for the time commitment if the panelists are employees (a common situation). For screening, the panelist must be known to have some normal functions of sensory acuity for the senses involved. Tests may involve discrimination of different products that have been spiked with different flavors, have different textures, or any combination as a function of different processing conditions. It is best to use products from the category they will actually evaluate after training. The persons do not have to be sensory superstars, but must have a roughly normal sense of taste and smell, for example. If a large panel is part of screening, the top scoring individuals may be invited for further service. They should also have good verbal ability, ability to describe their perceptions, be cooperative in a group situation, and not be afraid to voice their opinions. They need not be consumers of the product category, but they must have no objections to eating or drinking whatever the products are in the future tests. Typically, panel size is 8 to 12 individuals, but it is a good idea to have a reservoir of new potential trainees to replace persons who may drop out, leave the company, go on long-term leave of absence, and so on.

A.3.3 Methods

Most of the current descriptive methods are versions of the quantitative descriptive method outlined by Stone et al. (1974), although the historical antecedent was a method of flavor profiling (Caul, 1957). Once the recruits have been qualified, the training phase begins. There are two versions of training, which have been called “consensus training” and “ballot training.” In consensus training, the panelists taste a wide range of products from the product category and volunteer terms that describe the product’s appearance, aroma, flavor, texture, mouthfeel, and residual characteristics as applicable. The list is refined by eliminating redundant terms, vague terms, and those that refer to likes or dislikes. Reference standards may be found to illustrate good examples of products having those qualities, and these may be from other product categories or even be single chemical compounds in the case of flavors. Examples of reference standards can be found in the classic paper on the wine aroma wheel by Noble et al. (1987), and their use is discussed by Rainey (1986). The terms should only refer to one specific sensory attribute, and not be a combination such as “creaminess.” The early stages of panel training proceed with a lot of group discussion, rather than individual testing. Hence, the notion of a consensus is applied to the results of the training. Note that the training is simultaneously calibrating the attributes and concepts of the panelists, while building the eventual ballot for formal evaluations and data collection.

Once the terms are found, anchor words must be assigned to the low and high ends of each scale; for example, no sweetness to extremely strong sweetness. The anchor words need to be logical and chosen sensibly. In some forms of descriptive analysis, where there is a universal intensity scale, the intensity points are trained by example, and all flavor, taste, and aroma scales have the same references (usually on a 15-point or 150-point scale). For example, a "2" in intensity is exemplified by a 2% sucrose solution, among many other examples. See Meilgaard et al. (2006) for extensive lists of references for a universal intensity scale.

The second form of training is called ballot training, in which the terms are prespecified. Once again, panelists may taste a large variety of samples from the product category during training. Examples are shown to illustrate the different characteristics. In both methods, panelist consistency is checked during training, as well as panelist agreement. That is, panelists should be able to reproduce their own judgments on blind samples and panelists should agree within a certain range that the sample should be at a certain point on the scale. Panelists who are too low or too high may be offered additional training. The ins and outs of panel calibration are discussed in Chapter 7.

Once the panel is sufficiently "in tune," the actual evaluations may begin. Samples need not be presented to panelists in a group any longer, and they may fill out their ballots on an individual basis, as long as the product handling, preparation, and conditions of serving are identical. Samples are presented in random or counterbalanced orders, usually with three-digit random codes as identifiers. Scales are typically presented by computer screen and responses recorded by mouse clicking, but paper ballots may be used in some circumstances. An example of a ballot for peanut butter is shown in Figure A.2.

A.3.4 Statistical Analysis

The classic "bread and butter" technique for statistical processing of descriptive data is the analysis of variance (ANOVA; Lea et al., 1998; Naes et al., 2010). Generally, products and panelists are factors in the ANOVA model, and often replications. The specification of panelists as a factor allows the removal of inter-person variation from the error term. If all panelists see all products (usually an equal number of replicates) the analysis is what is referred to as a repeated-measures analysis, and/or a complete block design. These designs are highly desirable for reasons discussed in Chapter 9. After estimation of the panelist main effect, the remaining source of error variance is the product by panelist interaction, so this becomes the error term, or denominator of the F-ratio for products. This is because panelists are considered what is known as a random effect, or a random sample of all such possible panelists. This is in contrast to a fixed effect, where the treatment levels consist of specific values of some variable, like 2%, 4%, and 6% sucrose. The ANOVA models are different in that the construction of the error term is different, and so the practitioner must be careful to specify the correct model. For a case where products are a fixed effect and panelists are random (a common occurrence), this is sometimes referred to as a type III or "mixed" model.

Once a significant difference among products is found in the ANOVA, a second task begins. Now the mean values must be compared using variations of the t-test. Different techniques exist for this, most of which adjust for the fact that you are performing multiple paired comparisons, and thus the chance of a type I error is being inflated. Some are more conservative than others and require a larger difference between the means in order to be statistically significant. Common choices are the least significant difference test (LSD, very

PEANUT BUTTER DESCRIPTIVE BALLOT

1. Appearance

a) Coloration

□ □ □ □ □ □ □ □ □ □ □ □

very yellow very brown

b) speckles

□ □ □ □ □ □ □ □ □ □ □ □

none many

c) visible oil

□ □ □ □ □ □ □ □ □ □ □ □

none a lot

d) particles

□ □ □ □ □ □ □ □ □ □ □ □

smooth grainy

2. Aroma

e) fresh peanut

□ □ □ □ □ □ □ □ □ □ □ □

none very strong

f) roasted

□ □ □ □ □ □ □ □ □ □ □ □

none very strong

g) oxidized

□ □ □ □ □ □ □ □ □ □ □ □

none very strong

3. Flavor

h) sweet

□ □ □ □ □ □ □ □ □ □ □ □

very weak very strong

i) salty

□ □ □ □ □ □ □ □ □ □ □ □

very weak very strong

j) fresh peanut

□ □ □ □ □ □ □ □ □ □ □ □

very weak very strong

k) roasted

□ □ □ □ □ □ □ □ □ □ □ □

very weak very strong

Figure A.2 Sample descriptive analysis ballot for peanut butter.

4. Texture/mouthfeel

l) thickness

□ □ □ □ □ □ □ □ □ □ □ □

very thin very thick

m) oily

□ □ □ □ □ □ □ □ □ □ □ □

not oily very oily

n) adhesiveness

□ □ □ □ □ □ □ □ □ □ □ □

not sticky very sticky

o) melting rate

□ □ □ □ □ □ □ □ □ □ □ □

very slow very fast

p) drying

□ □ □ □ □ □ □ □ □ □ □ □

not drying very drying

5. Residual Characteristics

q) bitterness

□ □ □ □ □ □ □ □ □ □ □ □

none very strong

r) mouthcoating

□ □ □ □ □ □ □ □ □ □ □ □

not coating very coating

s) residual flavor

□ □ □ □ □ □ □ □ □ □ □ □

very weak very strong

Figure A.2 (*Cont'd*)

liberal), Duncan's multiple range test (or "studentized range," a good choice after a significant F-ratio), and the Tukey "honestly significant difference" (HSD). Most of these can be prespecified and chosen in statistical programs.

There is potentially a lot of information coming out of a descriptive analysis. So presentation of the results presents a challenge. A simple technique is to present the means in a table and use letter superscripts (typically a, b, c, etc.) to indicate significant differences. Products that do not differ share common superscripts. Those without an overlapping letter designation are significantly different. Other common displays include bar charts with error bars or letter superscripts, and spiderweb or radar plots, which produce a polygon for each product. In general, it is not necessary to include attributes that showed no significant difference among the products. Communicate the headlines. Also bear in mind that this kind of panel can become highly discriminating. Thus, a statistically significant difference is not necessarily a difference that is practically meaningful or important to consumers.

A.4 Affective Tests

A.4.1 Objectives

The general objective of an affective test is to find out the degree of consumer appeal of a product. This is the third major category of sensory testing. Difference tests can tell you if there is any change, descriptive tests can tell you how the product changed, and affective tests can tell you if it matters. Like other sensory tests, they are usually done on a blind basis; that is, with minimal concept information. This is in contrast to market research tests, which may be done with the full-blown concept being presented to the participants, in addition to a taste test. These tests can be divided into two general methods, one measuring acceptability of the products on a scale for liking/disliking and the second using a choice paradigm, where the best-liked product from a pair or group is indicated. The latter is a preference test. Affective tests are also called hedonic tests and are what most laypersons associate with a taste test. Because they generally use consumers of the product category, they are also called consumer tests.

There are several common scenarios for an affective test. One is the test of a new product prototype in a product development scenario. There may be a necessity to get a "read" on the new product and diagnose any potential problems before more expensive market research testing. The prototype may then go through further modifications before another consumer test. Minor product changes in an existing brand may need to be checked to make sure there are no objectionable sensory developments. These scenarios include ingredient substitutions and cost reductions. Shelf-life testing (stability testing) may involve some consumer evaluations, sometimes after a discrimination test or descriptive panel has found differences at a certain time point. Large-scale consumer tests may also be done with a finished product that is about to be introduced as a new product or a line extension to an existing product. Consumer testing is also required for advertising claim substantiation, and may be done against specific competitors in order to justify the claim. An example would be a statement like, "Taste America's best beef frank!"

These tests occur in different venues. Sometimes, users of the product are recruited from an employee pool to conduct a rapid in-house test. These people are less representative than a true consumer sample, but the turnaround time is quicker, and security is of course tighter. The second type of setting is in some central location, where consumers are recruited or invited to participate in a taste test in a sensory testing or market research facility. These are known as central location tests (CLTs). Many market research test agencies are available to do this kind of testing, and some have lists of consumers who have previously filled out product usage questionnaires. Thus, recruiting can be targeted and cost effective, although it is more of a convenience sample than anything you would call random. The CLT setting affords the tester good control over product preparation, serving, and eating conditions. The third scenario occurs when the product is sent home to the panelist. These are home-use tests (HUTs; also called IHUTs for in-home). These are, of course, more expensive than CLTs, but more realistic. The product may be used by all family members, and for an extended period. The questionnaire is usually administered at the end of the usage period.

A.4.2 Participants

The participants are screened for product usage. That is, they must be users of the product category. A category is a grouping of products that serve the same purpose, and often appear on the same part of the retail shelves. So cold breakfast cereals are a product category. If needed, the category may be further subdivided to narrow down the sample to members of the actual target market. So they may be screened to be users of pre-sweetened breakfast cereals, or even sweetened flaked cereals. Frequency of usage is an important consideration. If I only eat cold cereal once a year, I am probably not qualified to be in the test. The screening questionnaire may seek to find people with moderate to high usage of the product type. The questionnaire may also seek to find users of specific brands, either the company's own products or those of key competitors. In a CLT or HUT, the screening and recruitment process often seeks to have certain quotas of different age groups, genders, or other demographic variables, such as income groups. The typical group size is 100 or more consumers per product. In advertising claim substantiation, it is common to have 300 or more consumers per cell if there are different paired preference tests.

A.4.3 Methods

Consumer acceptability tests generally use the nine-point hedonic scale developed by the US Army. The scale points are shown in Table A.1. The phrases were chosen to represent approximately equal psychological distances, although the interior three are actually somewhat more closely spaced. They are generally assigned the numbers one (for dislike extremely) through nine (for like extremely) for purposes of data analysis. Consumers are asked to use this scale to give their overall opinion of the product. Consumers may also be asked to rate liking for appearance, flavor, texture, and so on, although their ability to understand and make distinctions about more specific attributes is questionable. But the consumer has an immediate and integrative reaction to a product, that could be called the "yum or yuck" response. Other related scales include satisfaction scales and appropriateness for a given situation.

Liking may also be probed concerning specific attributes using just-about-right (JAR) scales. These scales range from "not enough X" to "just about right" to "too much X." They give immediate direction about any needed product changes in order to optimize it. Of course, the consumer must understand the attribute in question, so this is limited to simple common terms that consumers can recognize clearly in the product. The goal is to make a product that has its data distribution centered on the JAR point or category, and is symmetric and peaked.

Preference tests are straightforward. Both products are tried by the consumer. Then they are asked to choose the product they liked the best. Sometimes a "no preference option" is allowed, but this complicates the analysis and is not favored by most sensory practitioners.

Table A.1 The nine-point hedonic scale

Assigned value	**Phrase**
9	Like extremely
8	Like very much
7	Like moderately
6	Like slightly
5	Neither like nor dislike
4	Dislike slightly
3	Dislike moderately
2	Dislike very much
1	Dislike extremely

However, the option may be required for some advertising claim substantiations. The orders of presentation, of course, are balanced across the participants, because there may be a preference bias for the first item. If more than two products are presented, the consumer may be asked to rank them from most liked to least liked. In another form, so called best–worst scaling, they are asked to choose the best liked and least liked. This test is still rare in food research.

Two major options are available for consumer test design. In monadic testing, each consumer only evaluates a single product. Usually more than one product is tested; thus, the design requires a different group of consumers for each product. In monadic sequential testing, the consumer receives one product at a time for evaluation, but is asked to evaluate more than one item. It is common with two or three products to have all consumers evaluate all products, but with more items, an incomplete block design can be used. A preference test is not possible with a monadic design, as there is nothing to compare with. If products are presented at the same time in a CLT, that is referred to as a side-by-side test. However, because we cannot taste more than one product at a time, it is simply another form of a sequential test, albeit with a short time interval. It does permit a direct visual comparison of the two items.

A.4.4 Statistical Analysis

For acceptability data gathered with the nine-point scale, parametric statistics are generally used, with *t*-tests for two items and ANOVA for more than two. After a significant ANOVA result, means may be compared by various post hoc tests, such as Fisher's LSD, Duncan, or Tukey tests. It is important, however, to look at the distribution of scores in the data set, for there may be pockets of consumers who like the item strongly and some who dislike the item strongly. Thus, the mean value can be misleading, and the researchers must be aware of the possibility of some consumer segmentation.

For the JAR data, a simple χ^2 test can be performed if the design is monadic; that is, there are different groups of people trying each product. Sometimes the data are collapsed into three categories for simplicity: below JAR, at or near JAR, above JAR. If the same persons try both products, as in a monadic sequential test, the Stuart–Maxwell statistic is appropriate, as discussed in Chapter 9. If hedonic scores are also gathered in the same study, it is possible to calculate the mean drop or penalty difference among persons who scored the product below or above JAR (as compared with the mean hedonic score for people at or near JAR for that product). This can give two important pieces of information. First, how many people are off-JAR for this product, and how potentially detrimental is that result in terms of the overall product appeal. Further information on JAR analysis can be found in the ASTM document edited by Rothman and Parker (2009).

The analysis of paired preference data is straightforward, as long as a choice is forced. Under a null hypothesis of equal preference for each product, a *Z*-score can be constructed from the normal approximation to the binomial, or an exact binomial probability can be computed using appropriate software. The *Z*-formula does not deviate much from the exact binomial probability, because the sample size is generally very large. The *Z*-formula is similar to eqn A.1, except that the test is now two-tailed and the critical *Z*-value is now 1.96. So in order to be statistically significant, the following inequality must be satisfied:

$$\frac{P_{\mathrm{w}} - 0.5}{0.5 / \sqrt{N}} \geq 1.96 \tag{A.3}$$

where P_w is the proportion for the winning product, the one gathering the most preference votes, and N is the number of consumers in the test. Sometimes the continuity correction ($1/2N$) is subtracted from the numerator, but this becomes negligible as the sample size gets large.

For tests with a no-preference option, various strategies may be applied to return the analysis to the binomial situation in order to apply eqn A.3. The votes may be ignored (thrown away), decreasing N accordingly. A conservative approach is to allot them equally to the two products, which of course dilutes the signal-to-noise ratio, as it follows the null hypothesis. Various other options are discussed in the body of this book, including Thurstonian analysis of the three-choice frequencies (see Chapters 4 and 8).

For ranking tests, various rank sum statistics can be applied, such as the Friedman analysis of variance on ranks and the Kramer rank sum test. Simple lookup tables for rank sum differences can be found in Lawless and Heymann (2010) and in the useful paper by Newell and MacFarlane (1987). Best–worst scaling is discussed in the paper by Jaeger et al. (2008).

A.5 Summary and Conclusions

Many years ago, Pangborn lamented three continuing problems in sensory evaluation. They were (1) lack of clear objectives, (2) a tendency to use one test method repeatedly regardless of the problem at hand, and (3) lack of proper selection of respondents in the test (Pangborn, 1979). Although these problems persist, training in sensory evaluation and adherence to good principles and practices will help avoid serious mistakes and useless data. The goal of this appendix was to give an overview of the basic methods used for different test objectives. The reader without formal training in applied sensory testing is encouraged to consult the texts mentioned in the opening and listed in the reference list below.

Many other types of testing are sometimes done under the umbrella of sensory evaluation. One example is threshold testing, which strives to determine the minimum amount of a substance that may be detectable by taste or smell, for example. However, many threshold tests use forced-choice procedures, and thus are just special cases of repeated discrimination testing. Another procedure is time–intensity scaling, in which the assessor indicates the intensity of a taste or flavor as it rises and falls once the food is tasted. The goal is to track the time course of the sensation, and identify temporal characteristics such as the total duration of the experience. But these are simply extensions of scaling, much like descriptive analysis. A third example is quality testing. But this involves some degree of difference scale, along with specification of the intensity of any key components and/or defects (Lawless & Heymann, 2010: chapter 17). Thus, it is a combination of a difference testing method and a descriptive method. So the fundamental three categories of testing (along with scaling, see Chapter 7) are still the best starting points for one's conceptual orientation to the field.

References

Bi, J. 2006. Sensory Discrimination Tests and Measurements. Blackwell Publishing, Ames, IA.

Caul, J.F. 1957. The profile method of flavor analysis. Advances in Food Research, 7, 1–40.

Chambers, E.C., IV, and Wolf, M.B. 1996. Sensory Testing Methods. Second edition. ASTM Manual Series MNL 26. ASTM, West Conshohocken, PA.

Jaeger, S.R.; Jørgensen, A.S., Aaslying, M.D., and Bredie, W.L.P. 2008. Best–worst scaling: an introduction and initial comparison with monadic rating for preference elicitation with food products. Food Quality and Preference, 19, 579–588.

Kemp, S.E., Hollowood, T., and Hort, J. 2009. Sensory Evaluation. A Practical Handbook. John Wiley & Sons, Ltd, Chichester, UK.

Lawless, H.T. and Heymann, H. 2010. Sensory Evaluation of Foods. Principles and Practices. Second edition. Springer Science + Business, New York, NY.

Lea, P., Naes, T., and Rødbotton, M. 1998. Analysis of Variance for Sensory Data. John Wiley & Sons, Ltd, Chichester, UK.

Meilgaard, M., Civille, G.V., and Carr, B.T. 2006. Sensory Evaluation Techniques. Third edition. CRC Press, Boca Raton, FL.

Naes, T., Brockoff, P.B., and Tomic, O. 2010. Statistics for Sensory and Consumer Science. John Wiley & Sons, Ltd, Chichester, UK.

Newell, G.J. and MacFarlane, J.D. 1987. Expanded tables for multiple comparison procedures in the analysis of ranked data. Journal of Food Science, 52, 1721–1725.

Noble, A.C., Arnold, R.A., Buechsenstein, J., Leach, E.J., Schmidt, J.O., and Stern, P.M. 1987. Modification of a standardized system of wine aroma terminology. American Journal of Enology and Viticulture, 38(2), 143–146.

Pangborn, R.M. 1979. Physiological and psychological misadventures in sensory measurement or the crocodiles are coming. In: Sensory Evaluation Methods for the Practicing Food Technologists. M.R. Johnson (Ed.). Institute of Food Technologists, Chicago, IL.

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

Rothman, L. and Parker, M.J. 2009. Just-About-Right Scales: Design, Usage, Benefits, and Risks. ASTM Manual MNL63. ASTM International, Conshohocken, PA.

Stone, H., Sidel, J., Oliver, S., Woolsey, A., and Singleton, R.C. 1974. Sensory evaluation by quantitative descriptive analysis. Food Technology 28(1), 24, 26, 28, 29, 32, 34.

Stone, H., Bleibaum, R., Sidel, J., and Thomas, H. 2004. Sensory Evaluation Practices. Third edition. Elsevier/Academic Press, Amsterdam.