Reading: Jackson pp 564-590

Sensory Evaluation Basics

by Harry T. Lawless

The overriding principle of sensory evaluation is to match the sensory technique with the problem at hand. This requires a logical decision-tree approach to test design. Most questions about perception of flavors or products will fall into three categories. First, people want to know, "Are these two products different?" This calls for the overall difference test, also referred to as a discrimination test. These tests usually take the form of a forced-choice procedure, where participants are asked to select one choice from among a set of products in which only one is physically different from some standard sample. The second common question is, "How are they different?" In other words, the goal is to specify, in perceptual terms, how products differ, in what qualities have they changed and to what extent. This set of procedures is referred to as descriptive analysis. In its most common form, a group of trained individuals examines the products and provides numerical ratings for the perceive intensity of each attribute. This provides quantitative sensory specification of each product that may be compared statistically. The third common question concerns consumer likes and dislikes. These tests are generally conducted with untrained persons who are usually users or purchasers of the product. They are asked to provide quantitative ratings describing the strength of their liking or disliking for the product as a whole, and may also be probed about their opinion regarding specific characteristics. Alternatively, they may simply be asked to pick which product they like best from a set of alternatives.

In historical sensory analysis of some standard commodities, the descriptive approach and acceptability testing were combined in the "quality judgments" of expert tasters. The experts were able to identify defects, judge their severity, and produce quality scores that would presumably reflect consumers' rejection of substandard products. This approach has limited applicability to newly developed and processed or engineered foods, where standards for quality are not yet defined. It is also problematic for products in which sensory segments exist, i.e. consumer groups that have different yet specific profiles of what they like in a certain product (tastes great? less filling?).

Since these methods involve a controlled stimulus and response measurement scenario using human participants, sensory evaluation borrows some practices from the behavioral sciences. In order to minimize biases that may affect the validity or accuracy of a test, blind coding and control of presentation order are critical. "Blind" coding is usually achieved by labeling each sample with a meaningless name such as a randomly chosen three digit number. Participants are provided with only enough information about the sample to insure that it is viewed in an appropriate frame of reference or category. Controlling stimulus order may be achieved by fully counterbalancing orders or by using a design such as a Latin square, which would place each product in each position an equal number of times, when viewed across all participants. Alternatively, order might be fully randomized if the number of observations is large enough. Other critical items to control are all the physical variable that would be expected to influence sensory impact: Concentrations, volume, temperature, and so on. These concerns seem second nature or even old hat to behavioral scientists, but they can easily be overlooked in applied situations.

Discrimination testing.

A discrimination test is called for when the objective is to determine whether any difference is perceived between two products. The nature of the difference is usually not specified -- it is up to the test participants to see if they can find a point of difference. Since a finding of no difference may have important business implications, failure to reject the null is an actionable outcome in these tests. Thus the power and sensitivity of the test is important, and beta risk is an important consideration.

If the difference is studied as a function of different levels (systematically varied) of some ingredient, the experiment resembles a measurement of difference thresholds. For example, the determination of a just-noticeable difference is closely analogous to the discrimination test objective, when several products with different levels of a flavor are compared to some control. This is logically related to historical psychophysical methods such as the constant stimulus method.

Variations.
The most common forms of the discrimination procedure are the triangle test and the duo-trio procedure. The triangle test is a three-alternative test in which one sample is different from the other two. The test is counterbalanced for the identity of the odd sample (both ABB and BAA used) and its position in tasting (ABB, BAB, BBA). Chance performance is one third, and performance in a group above that level provides evidence for a perceivable difference. This differs from traditional forced-choice procedures in which subjects would be directed to choose the strongest or weakest stimulus. Foods are necessarily multivariate. Since it entails comparisons of similarities (or differences), rather than simple intensities, the triangle test is more difficult than choosing the strongest of three samples. In the duo-trio test, a sample designated as a standard or control sample is presented for inspection. Then two samples are presented, and the participant is asked which of the two matches the control. Chance performance is 50%. In neither test is are the participants directed to any specific sensory attribute -- the test is for the existence of any difference whatsoever.

Participants in these tests should be screened for minimal acuity in discriminating differences in the products or sensory modalities to be tested. Since the tests are often used as a first step in a sequence of tests, there is little attempt to make the selection of participants be representative of consumers as a whole. Rather, discriminative ability is key. The tests are generally conducted in a laboratory environment with control over sample preparation, temperature, lighting, noise, etc. If no difference is found under such conditions, logic dictates that most consumers would not notice a difference in less controlled situations. However, this logic is not airtight. Consumers have multiple opportunities to interact with a product once it enters the home, and to solicit opinions from other family members. There is always the possibility that what goes unnoticed in a short taste test might be detected once familiarity with the product is gained.

Forced-choice procedures are also used on some occasions. In these tests, the subject is directed to a specific attribute, e.g. "choose the sample that is sweeter from this pair." Such tests are designates as n-alternative forced choice tests (2-AFC for a paired test , 3-AFC for a test with one target and two controls, and so on). There are both theoretical and empirical reasons to believe that they are more sensitive than the overall difference tests. Since the participant's attention is being directed to a sensory attribute which is expected to differ, detection of the difference should be enhanced for attentional reasons. Conversely, overall difference tests such as the triangle may lose sensitivity since they entail the risk that some participants may focus on differences which are artifactual, a function of serving order or even of their momentary state of adaptation, and fail to attend to the attribute or attributes which are expected to differ systematically as a function of the ingredient change.Sensory evaluation classes asked to perform various discrimination tests on a 10% difference in sucrose concentration always find that the paired test with directed attention was usually much more sensitive on a chance-corrected basis and more likely to achieve statistical significance.

Descriptive analysis.

The most generally useful and highly informative class of sensory tests are the descriptive analyses. These techniques attempt to provide a quantitative specification of all the sensory attributes of a food or product. This is typically achieved using a set of scales, each of which provides a numerical response for the perceived intensity of a given attribute. Each sensory attribute represents a (presumably) independent and elemental sensory experience. The results are useful for specifying sensory changes in product development as a function of ingredient, packaging or processing variables and for shelf-life and quality control questions. The data are also used for correlation with consumer judgment for purposes of building predictive or explanatory models of factors driving likes and dislikes. Since they are quantitative and analytic in nature, the sensory specifications are also sometimes examined for correlation with instrumental measures of food properties.

Variations.
The earliest method for descriptive analysis was the Flavor Profile method. A group of extensively trained panelists would make judgments about the perceived intensity of all the flavor components of a product, in the order of their appearance. The individual profiles would then be discussed, and a consensus profile was put together under the direction of a panel leader. While this procedure was a great improvement over the liabilities inherent in using a single expert taster, further advancements were possible. A technique called Quantitative Descriptive Analysis (QDA) brought aspects of behavioral testing methodology to the descriptive test . The simple category scale used in Flavor Profile was replaced with an unstructured (presumably less biasing) line scale, anchored with suitable words at the low and high ends. More importantly, this technique was amenable to experimental design and statistical analysis. Replication was a standard feature of the design and allowed for evaluation of test reliability. Analysis of variance became the routine statistical procedure for these data. Repeated measures analysis could be applied to partition judge effects from product differences. Statistical measures of central tendency rather than consensus values became the framework of the "profile" and statistical significance testing provided the criteria for the existence of product differences.

Terminology issues.
A major hurdle in the development of a good descriptive analysis is the selection of useful terms. While there is some agreement about four basic tastes, other points of view add other taste qualities, such as the umami taste of monosodium glutamate. The realm of olfactory characteristics and texture words are less uniformly agreed upon. Early systems for expert grading of foods centered on physical sources of defects, rather than perceptual description of subjective experiences ("oxidized flavor" as opposed to "blue color"). Such systems are problematic in that many different sensory qualities may be subsumed under a single physical defect, e.g. cardboard-like, painty, fishy and tallowy notes may all arise from oxidation flavors. Wine tasters sometimes discuss "reduction flavors" such as the mercaptans which arise from microbial degradation. However, "reduction" may lead to a host of different putrid sulfurous odors, so the specificity of this term is lacking. Recent publications have focussed on this linguistic problem, and criteria for the practical utility of sensory attributes have emerged. These criteria include simplicity, lack of redundancy with other terms and the feasibility of finding a physical reference standard to serve as a training example. Measurement criteria such as reliability (precision) and validity (accuracy) may also be brought to bear.

Training panels of individuals to think about and to use words in a similar way is obviously a major hurdle for descriptive techniques. Many consumers confuse sourness and bitterness, but this confusion is easily rectified with examples, e.g. citric acid for sour and caffeine and quinine for bitter. Training panels to recognize volatile aromatics, (aromas and flavors) is more difficult, but is also aided by use of examples or reference standards. These examples help to categorize and calibrate the qualitative perceptual space for trainees. Ó Mahony and coworkers have framed the process as one of concept learning or in their terms, concept alignment.

Some practitioners have carried the process a step further, and not only calibrated the panels with qualitative references, but then attempt to calibrate the psychophysical intensity curve as well, giving panels examples of low and high levels of each attribute. This kind of intensity anchoring was a critical aspect of the original texture profile method. For example, reference standard for perceived hardness ranged from low end anchors like boiled egg white to high end anchors like peanut brittle. Some descriptive analysis practitioners carried this approach over into flavor work, and even went so far as to suggest cross-modal scaling, i.e. a common perceptual intensity scale for tastes and flavors. Whether panelists can be so fully calibrated remains a source of some controversy. For example, subjects who eat diets high in red pepper compounds such as capsaicin, become chronically desensitized to those flavors. It would seem fruitless to force them into a perceptual intensity scale on the same basis as non-consumers of hot spicy cuisines, who are more sensitive.

In common practice, the profiles are represented by a line graph in polar coordinates, with the attributes forming equally spaced rays (arrangement otherwise arbitrary) and distances from the origin along each ray representing the mean value for a product on that ray. The points are then connected, forming a spider-web or polygon with a sometimes characteristic and recognizable shape for the control product. This is thought to aid in recognition of product differences due to the human ability to perceive shape, where the differences would be more obscure in a bar graph.

Affective tests.

The third important type of sensory test involves questioning consumer likes and dislikes. This question is phrased in two ways. One may ask about the liking or disliking for a product, perhaps a single product without reference to another product for comparison. Generally, these data are collected as ratings on a numerical scale, such as the balanced nine point category scale introduced by the food research section of the U. S. Army Quartermaster Corps in the 1950s. This scale runs from "dislike extremely" to "like extremely" with a neutral category at the center of the scale. This degree of absolute liking and disliking should be referred to as "acceptability."

The second way to phrase the question is in a choice or ranking between two or more products, usually a paired test to see which is liked better. This should be referred to as "preference," although there is widespread misuse of this work in the literature to refer to rated acceptability. Preference or choice data are usually analyzed by means of binomial distribution statistics, since discrete outcomes are counted (numbers of people preferring one item over the other, as a proportion of the total). It is widely believed that preference data are more sensitive than rated acceptability, since two products can get the same acceptability rating on a category scale, but there might be a slight preference for one over the other. Empirical data supporting this belief are not found in the scientific literature. Furthermore, preference tests tell little or nothing about the overall level of acceptance, since one product might be preferred over the other, but both might be unacceptable. On the other hand, acceptability ratings can provide information on the direction of presumed preference.

The emphasis in both procedures is on obtaining a representative sample of consumers for the test. Several principles apply. First, laboratory personnel generally make poor choices for participants. They come with a technical and potentially biased frame of reference for evaluating the products. If consumers are recruited, they should be regular users or purchasers of the product. Finally, since the variability in personal preferences is usually quite high, large numbers of participants are usually needed (N> 100, as a rough rule of thumb). With such larger samples, it is possible to look for segments or groups of consumers with different preference patterns, rather than simply looking at overall means for different products or other measures of central tendency.