When People Are The Instrument: Sensory Evaluation Methods

by Irene Gengler, Sensory Testing Service Inc.

http://sensorytestingservice.com/index.html

Abstract:
Sensory evaluation is a field that measures product attributes perceived by the human senses. The inherent variability of human responses has led to special methods and procedures for their measurement. Understanding the type of response being measured is important for designing research. Different methods have been developed, and by understanding the core principles of usage for these methods, one can improve the quality of the human measurement. The data then becomes more useful for developing and maintaining successful products.

Sensory Evaluation

Measurements using people as the instruments are sometimes necessary. The food industry had the first need to develop this measurement tool as the sensory characteristics of flavor and texture were obvious attributes that couldn't be measured easily by instruments. Starting in the 1940's, the first trained panels were developed in an effort to make measurements of food more objective, given the inherent subjectivity and variability of human evaluators. Eventually the field of Sensory Evaluation emerged, and was applied to a variety of other product types. The following definition written by the Sensory Division of the Institute of Food Technologists has held up well over the years.

> ***"A scientific discipline used to evoke, measure, analyze and interpret reactions to those characteristics of foods and materials as they are perceived by the senses of sight, smell, taste, touch and hearing."***
> -Institute of Food Technologists (IFT), Sensory Division

This applies to a range of products, including cosmetics, household cleaners, paper products, fabrics, tobacco products, pharmaceuticals, automobiles, etc., with more applications all the time. Anything that has sensory characteristics perceived by one or more of the human senses can be measured. The question is how to design methods that deliver consistent, reproducible results.

Methods

A. Descriptive Analysis

The first attempts to use people as measurement tools were made with trained panels that measured the intensities of sensations from food samples without the like or dislike response. For example, saltiness was rated for intensity only, not how well it was liked. Other more complicated attributes, like "caramel flavor" or "cohesive" texture required training panelists so that they were all describing the same thing consistently. Various ways of training panelists have been developed, and the methods are generally referred to as **Descriptive Analysis**. It is the most analytical method, and describes attribute intensities without assessing liking for them.

B. Acceptance Tests

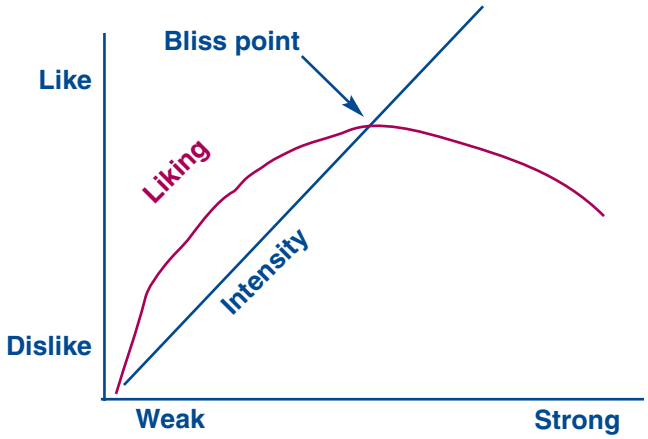
The other human response to products is, of course, liking or acceptability. These tests are usually referred to as **Acceptance Tests**, and are best done with a large group of respondents because of the subjective nature of the response. The general population can vary greatly in product preferences, so it is important to use respondents representative of product users or the target market. Non-users could easily provide different results that would mislead product decisions. In acceptance tests, validity is the primary issue as it is predictive of marketplace success.

As an example of how these two methods differ, below is a plot measuring ***intensity*** of a sensation versus ***liking*** of that sensation:

Continued on page 9

When People Are The Instrument: Sensory Evaluation Methods

Continued from page 8



The plot shows how intensity of an attribute, e.g. saltiness, can increase with higher concentrations, while liking for the attribute reaches a peak and then declines. This general pattern occurs quite commonly, although the pattern varies with the attribute and its interaction with other attributes. Measurement of the intensity response is best done with small trained panels using Descriptive Analysis. These panels measure specific sensations, without emotion (no like/dislike), and make them more objective and reproducible. Descriptive Analysis panels are typically 8-10 respondents, so they are too small a group for measuring liking and potentially unrepresentative of the target market. This leads to the need for both methods in many situations.

The liking response is more variable than intensity, because it is an emotional response based on a variety of factors. Acceptance tests done with appropriate respondents provide liking scores for the products, and related information. Marketing strategies can also be explored in this type of research. Since the respondents are not trained, they have limitations in describing specific sensations.

C. Discrimination Tests

The third broad category of sensory tests is **Discrimination Tests**. Often referred to as difference tests, they are designed to measure the likelihood that two products are perceptibly different. One of the most common types is the Triangle test, in which the evaluator receives three samples, among which two are the same and one is different. The task is to identify the different sample. Evaluators perform best if they are familiar with the type of test, and might be trained panelists. Responses from the evaluators are tallied for correctness, and statistically analyzed to see if there are more correct than would be expected due to chance alone. This test is generally best as a screening tool, because of the high risk that samples may be slightly different when a no difference result is found. This leads to an important point for sensory testing:

Different test methods answer different questions, so you must be clear about what question(s) you are asking.

Sensory Questions

Questions about products are what lead to the need for testing that measures human responses. For example, you may be asking:

- How is your product different from others in the marketplace?
- Is the latest formulation different from the last one?
- How do formula and processing changes affect the product?
- What changes occur in the product as it ages?
- What are the likes and dislikes for the product?
- Will the product user remember the product's sensory characteristics?

Continued on page 10

When People Are The Instrument: Sensory Evaluation Methods

Continued from page 9

These are just a few questions that come up frequently, many others occur when circumstances change.

Method Selection

Determining the type of test to use can be difficult, as one may tend to choose the familiar or convenient method. Having clear objectives based on the questions you are trying to answer should dictate the test method. Often you have more than one question you need to answer, and you must make choices about methods based on your resources and time. Often both an Acceptance test and a Descriptive Analysis panel are required. Acceptance tests measure **liking** for test products, while Descriptive Analysis panels more precisely measure the product attribute **intensities**. Despite attempts to collect both intensity and liking on specific attributes, they are usually best measured separately. Using the wrong method can give you information, but may not answer your question.

Question	Method	Respondents	Sample Size (N)
Are they different?	Triangle, Duo-trio, PC, rank, sort	Test-wise, possibly trained	30+ (or 15 with a replicate)
What is the diference?	Descriptive Analysis	Sensitive, screened, trained	8-12 panelists, replicates
How are they liked?	Acceptance: Central Location Test (CLT) Home Use Test (HUT)	Represent end user	30-100+ users

Product Evaluation Procedures

Test samples for human evaluation require special attention to minimize the many biases that can occur. Blind codes, such as 3 digit numbers, are used to mask sample identity. The order (sequence) that samples are evaluated is balanced across respondents so each sample is evaluated 1st, 2nd, 3rd …. Nth as equally as possible. Order bias cannot be eliminated, but can be blocked across respondents to reduce its effects. Score sheets should be designed to facilitate ease of evaluation and not be too long. Both physiological and psychological fatigue is considered in the sample evaluation protocol, and enough time between samples for recovery of the senses. Ratings for samples are influenced by the context in which they are evaluated, and scores are relative to the other samples in the test. Because scores are not absolutes, a control or reference sample may need to be included for data interpretation.

Rating Scales

Acceptance tests usually involve category scales, most commonly Hedonic (liking) scales that are 9 points in length. The midpoint is neutral, and the other points reflect increasing or decreasing degrees of like or dislike. There are many variations on the Hedonic scale, but the classic 9-point scale has seen the most use. Considering the need for the distance between points to be perceived as equivalent, the words under each were researched to make them as equally spaced as possible.

☐ Dislike Extremely	☐ Dislike Very Much	☐ Dislike Moderately	☐ Dislike Slightly	☐ Neither Like nor Dislike	☐ Like Slightly	☐ Like Moderately	☐ Like Very Much	☐ Like Extremely

In most cases **Descriptive Analysis** panels use graphic line scales to rate intensities so that the panelists are not limited to discrete points. These scales can increase discrimination among samples, and are usually preferred by the panelists. They require later conversion to numerical measurements via manual or scanning entry of the results if direct computer entry is not available.

Continued on page 11

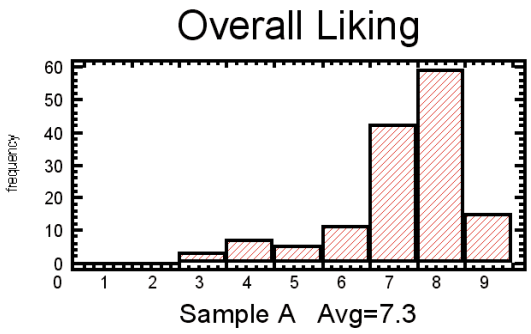
When People Are The Instrument: Sensory Evaluation Methods

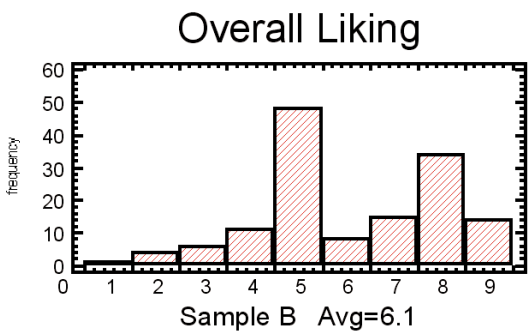
Continued from page 10

Data Analysis

There are many ways to analyze data, and sensory results can be particularly challenging due to people as the measurement tool. Test respondents don't always perform as expected, or participate when needed, leading to missing data. **Understanding how responses change due to product differences versus change due to panelist differences is important**. Replication is a good way to assess Descriptive Analysis results to evaluate panelists for consistency and agreement with other panelists. Acceptance tests usually have disagreement between respondents, and the challenge may be to decide if they represent different market segments (see histograms below). Scale usage in acceptance tests varies from panelist to panelist, and balanced block designs where each respondent sees all products help to minimize this difference. Significance of results from Discrimination tests can be determined from tables of probabilities based on sample size and number of correct responses.

The use of parametric statistics for rating scale data has been debated, since the criteria for usage (normally distributed data) is not always met. In most cases the benefits of using parametric vs. non-parametric analysis outweigh the disadvantages. Examining the distributions and variance in the data is always important and might indicate reason to use nonparametric methods.





For most tests involving rating scales, Analysis of Variance (ANOVA) is used for analyzing the results, followed by post hoc tests for means separation of specific samples. e.g. Duncan's, SNK, LSD, etc. Correlation and regression between attributes may also be done, and sometimes it is helpful to apply data reduction techniques such as Principle Component Analysis (PCA), Factor Analysis, etc.

Designed experiments can be useful for generating sample sets to understand technical variables that affect sensory attributes, and allow for more statistically powerful analysis of the results. The most complicated analysis involves regression to relate liking scores to Descriptive Analysis. A sample set that varies on the important sensory attributes is then necessary to understand key drivers of liking and build a model that is predictive for future use.

Applications

A few applications for sensory evaluation are detailed in the table on the next page:

Continued on page 12

When People Are The Instrument: Sensory Evaluation Methods

Continued from page 11

USAGE	EXAMPLES	TEST METHODS	DATA ANALYSIS
Measure product attribute intensities and their effects on other materials	Shampoo with *low density foam* adds perceived *moistness* to hair; *darker* color of coffee increases perceived *bitterness*	Descriptive Analysis	ANOVA Correlation/ Regression DOE
Measure sensory distance from a target sample	Low fat formula for cream cheese has *less mouth coating* than full fat cream cheese; *readability* of ebook is lower than printed page	Descriptive Analysis	ANOVA Correlation/ Regression PCA, Factor Analysis
Determine sensory effects of technical variables, e.g. formula and/or process changes	Experimental design identifies variable interaction that increases *strength* in paper towels; lower cost formula for detergent identified with equal acceptability	Descriptive Analysis Acceptance Tests	ANOVA Correlation/ Regression DOE
Compare competitive products for sensory differences and acceptability	Generic facial tissues have *lower smoothness* compared to market leader; sports shoes vary on *arch support* and *ease of lacing*	Descriptive Analysis Acceptance Tests	ANOVA Correlation/ Regression PCA, Factor Analysis
Monitor sensory changes during product storage	*Rancid flavor* in peanuts affected by humidity, lotion stored at high temperatures becomes *thinner,* toothpaste loses *flavor intensity* in some packages	Descriptive Analysis Discrimination Tests Acceptance Tests	ANOVA Correlations DOE Probability Tables (difference tests)
Quality control of sensory attributes in plant production	Pastries with *reduced flakiness* identified and reworked; tape with *low adhesion* discarded before shipment	Descriptive Analysis Discrimination Tests	ANOVA Probability Tables (difference tests)
Estimate likelihood two formulas are perceptibly different	The current formula has been cost reduced or an alternate ingredient/supplier is needed	Discrimination Tests	Probability Tables
Compare acceptability of proposed versus current formula	The current formula has been cost reduced or an alternate ingredient or supplier is needed	Acceptance Tests	ANOVA

Benefits

Human measurements are variable, but can be made more reliable if appropriate methods and procedures are used. As with any testing, resources are needed for good measurements. Sensory data can facilitate good decisions on a variety of issues, and the improvement in the quality of the information collected will have long-term value for product decisions.

Summary

The field of sensory evaluation is relatively new, and requires a variety of skill sets ranging from science, mathematics, business, physiology and psychology. While Food Science was the first area to use sensory methods extensively, they have spread widely to other products that have attributes perceived by the senses. Methods to improve the measurement of human responses have been developed and keep getting better. Different methods provide different types of information, so it is important to identify the question you are asking to design and execute testing that yields the human measurements you need.

References

"Sensory Evaluation Practices" by Herbert Stone and Joel L. Sidel, 3rd edition 2004 Academic Press

"Sensory Evaluation of Foods: Principles and Practices" by Hildegarde Heymann and Harry T. Lawless 1999 Chapman & Hall

ASTM E-18 Committee on Sensory Evaluation: General principles and specific product guides.

Society of Sensory Professionals – www.SensorySociety.org