Laboratory Flavor Panels

By
Mrs. Lois A. Sather
Assistant Food Technologist
In Charge, Flavorium
Department of Food and Dairy Technology
Oregon State College

For the purposes of this paper, flavor will be considered as the over-all sensation perceived in the mouth when food or beverages are consumed. Flavor is a complex of stimulations of which the three main components are aroma, taste, and texture or feel.

Laboratory flavor tests may be divided into two main categories, difference and preference tests. A third type of test may be listed but is not included under laboratory tests; this is consumer acceptance testing. It is impossible to determine over-all consumer acceptance in the laboratory because of the many other factors besides flavor which are involved in this type of test such as price, convenience of serving, advertising, etc. Over-all consumer acceptance can be determined by market and use tests.

Difference tests are concerned only with judging or rating of the food itself, not with the judge's like or dislike for the food. In the laboratory, difference tests may be employed as a tool in quality control to measure such factors as uniformity of product. Difference tests may measure differences or degree of difference in flavor, or the intensity or degree of specific quality factors related to flavor, such as sweetness, acidity, or saltiness.

Preference tests are those tests in which the judge's like or dislike for a food product is recorded. In other words, the judge's reaction to the food is being measured as opposed to scoring the food itself as is done

in the difference test. Preference tests include statements as to a judge's preference, acceptance, like, dislike, which may be expressed in terms such as excellent, fair, poor. Preference and difference tests are usually comparative while acceptance tests are usually absolute.

The use of laboratory flavor panels involves a careful consideration and awareness of the psychological and environmental factors which influence a person's flavor opinions. It might be said that everything a person is or does influences his flavor opinions. These influencing factors may be separated into two classifications according to the degree of control usually exercised in a laboratory situation.

Influencing factors which are normally not controlled but should be recognized and considered when interpreting the results are:

1. The judge's home environment, economic status, nationality, religion, age, sex, health, etc.

2. Weather conditions such as temperature and humidity.

Factors which should be controlled or influenced are:

1. The place of testing. Use of an open laboratory or of booths. The immediate environmental conditions of the test area, such as ventilation and light.

2. Sample selection, method of serving, coding, wording of the ballot, etc.

3. The judge's mental attitude or interest in the tests.

In conducting laboratory flavor tests, the first factor to consider is the samples to be judged. Are the samples truly representative? This seems very obvious but should not be taken for granted. The results of any flavor tests are only as valid as the samples will permit. The sampling method should be designed to insure representative samples. In an in-plant situation with biased personnel, the sampling method is sometimes questionable. A disregard for the importance of or lack of understanding of flavor testing tech-

niques by other cooperating personnel or departments may contribute to selecting non-representative samples. The administrator of the flavor panel should know the complete history and handling of each sample to be sure that no variables other than those being tested have influenced the sample.

The flavor testing area should, if possible, be separated from the preparation area so that the judges can concentrate and cannot observe the samples being prepared. This may be achieved by judging in a room adjacent to the laboratory or by means of a portable screen to separate off a section of the laboratory. Screening off a section of the laboratory is not too satisfactory if aromas from food preparation cannot be controlled.

The judges should be seated at a table or counter in such a manner that they cannot observe or watch any other judge's reaction or scoring of the product. Inexpensive partitions can easily be constructed to fit any counter or table and achieve an individual booth situation for each judge.

During judging, there should be no verbal comments or discussion of the samples. When training judges for <u>difference</u> testing, it may be desirable to have a round table discussion of the samples before and after judging in order to teach the judges to recognize and score certain quality factors on a common basis.

Controlled and uniform preparation and serving of the samples to the judges is essential. The samples should be at the same temperature, plus or minus two degrees, and this temperature should be the temperature at which the foods are normally consumed. Consideration should be given to the fact that at very high temperatures, 120°F. and above, and at low temperatures, 40°F. and under, taste buds are less sensitive and full flavor perception is not possible.

Foods of small unit size such as peas, corn, raspberries, etc., need no size preparation. Products of larger unit sizes, such as canned peach or pear halves, should be diced into smaller unit pieces and carefully mixed so as to secure a uniform sample. Grinding or macerating is not desirable as it incorporates air and alters the normal flavor of the product.

If paper cups or plates are used in serving, care should be taken to be sure the paper is not contributing a flavor of its own to the product. The serving container should also be adaptable to legible coding.

The samples should be coded in such a manner that the judges cannot distinguish the samples by the codes nor be subjected to code bias. For example, if the samples are numbered 1, 2, 3, or lettered A, B, C, a coding bias could be caused because people associate 1 or A with "first" or "best" and might tend to score this sample higher. In the flavorium, three-digit random numbers are used for coding the samples. As the flavorium is equipped with six flavor judging booths, six three-digit random numbers are assigned to each sample so that if judges did try to discuss the samples during testing, they would not have the same sample numbers to compare.

When multiple samples are being judged at a given time, the samples should be presented in random order to the judges in order to cancel out bias by order of testing. Some judges tend to show a bias for the first sample tested while other judges tend to show a bias for the last sample tested.

The number of samples which may be tested at one time is debatable. There is no set rule which can be given. Experience has indicated that this might vary with each product according to the intensity or type of flavor present and also vary with the judges as to their capacity or interest. In a mild flavored food, such as pears, it may be possible to judge as many as

6 or 8 or even more samples at one time. In a more intense flavored product, such as peppermint oil, judging of only one or maybe two samples may be possible at one time. Another factor which should be considered when deciding on the number of samples is the influence of one sample on the other. If a mediocre sample is scored with a poor sample, it will tend to score higher than when scored with a very good sample.

People who serve as judges are vital to any flavor test. The four questions most often asked concerning judges are:

1. Who should serve as judges?
2. How many judges are necessary?
3. Should the judges be trained or selected
4. How much information concerning the test should the judges be told?

The answer to these questions depends upon whether the test is a difference or a preference test. For flavor difference testing, it may or may not be necessary to train the judges. If the detection of very minute or small changes in flavor is desired, then a high degree of training of the judges may be necessary. Probably the best known example of judges who are highly trained to detect small differences are the professional tea, coffee, and wine tasters. However, in most laboratory situations, the degree of flavor differences being evaluated is not this minute. If a large number of people is available from which to choose judges, screening tests to choose those people most capable of consistently detecting the differences being tested might be all that is necessary and actual training would be unnecessary.

The number of judges for difference testing may be relatively small depending upon the number of replications which are made and also on the magnitude of the differences to be determined. In the flavorium, for the degree of flavor differences with which we are usually concerned, ten to fifteen judges and four replications are usually adequate to determine if

the judges as a group can detect a flavor difference. For 20 to 30 judges, two or three replications may be adequate.

For preference testing, it is undesirable to train or screen the judges. Even those people working with the development or preparation of the samples or in any other way closely associated with the samples should not be included in the preference panel because of the possibility that prejudices may have developed. Coding the samples does not always eliminate this factor as these judges may recognize certain flavor characteristics because of their association with the product and thus be prejudiced in their scoring.

As preference tests are for the purpose of securing an indication of possible consumer acceptance, the number of potential judges is almost unlimited. Enough judges should be used to be representative of the population of consumers with which the food product is concerned. The number of judges used will vary depending on the type of population, sampling design, cost, and method of testing. In the flavorium, a minimum of 100 college students serve on the preference panels. However, on consumer preference panels conducted by commercial concerns before marketing a new food product, the number of families may vary upwards from 100 to as many as 10,000 families. In preference tests, the main danger is in drawing an unwarranted conclusion because too few judges or nonrepresentative judges were used.

The morale or mental attitude of the judges is a very important factor which should never be neglected by the administrator of the tests. If judges are interrupted during their usual work, they resent the intrusion and this is reflected in their often hurried opinions and makes the accuracy of their judgments questionable. It is desirable to discuss with the judges as a group the best time of day for testing and then to send a written memorandum to each judge so that he can budget his time accordingly. In

the flavorium, when using personnel from the Food Technology Building, it has been found that 9:45 a.m. and 2:45 p.m. are the most suitable times as far as the judges are concerned as this is just preceding the 10:00 a.m. and 3:00 p.m. coffee breaks when they would usually be interrupting their work. When using student judges, they are allowed to test at any time which is most convenient for them between 8:30 and 11:30 a.m. and 1:30 to 4:30 p.m. After testing, coffee, punch, cookies or donuts, or what we classify as "treats" are made available to them.

In order to maintain the judges' interest, it is desirable to tell them certain basic information concerning the importance and the need for the tests to be conducted. The difficulty is in giving the judges enough information to maintain their interest in the tests without giving them too much information which might tend to influence or prejudice their scoring. As each situation varies, no set rule can be given concerning what information should be given the judges. However, preconsideration and analysis should be made of all information as to its possible effect on the judges' scoring of the samples.

In replicated small panel difference tests, the judges are naturally curious about, as they put it, "How did I do?", or "Did I get them right?". As far as possible, their questions should be answered. It is sometimes possible, depending on the nature of the tests, to post the previous day's scoring so that each judge can compare his scoring with the panel as a whole. This may also increase the accuracy of the panel as well as help to maintain interest.

There are several well-known methods of difference testing which may be used in the laboratory. It is impossible to discuss all of these but five of the most widely used procedures are:

1. Triangular Test.

In this test, three samples are presented to the judge. Two of the samples are identical and one sample is different. The judge is asked to indicate the duplicate sample or to indicate the odd sample. The triangular test is very time-consuming when a large number of samples is to be judged and it does not give the degree of flavor difference other than is indicated by the number of correct judgments. Rapid analysis of the results is possible as tables have been published which give the number of correct separations necessary in relation to the total judgments for statistical significance at the 1% and 5% levels.

2. Duo-trio Test.

In the duo-trio test, three samples are presented to the judge. One of these samples is labeled as a reference or standard sample and the judge is to indicate which of the other two samples is the same as the reference sample or which is different from the reference sample. The number of correct separations required for significance is greater in the duo-trio test than in the triangular test since the judge has a 50% chance of guessing the correct sample in the duo-trio test, whereas, in the triangular test, there would be only a 1/3 chance of a correct guess. The duo-trio test is also time consuming when a large number of samples is to be judged and gives degree of difference only be inference from the total number of correct separations. The analysis of results is the same type as in the triangular test and tables have been published for this purpose.

3. Multiple Comparisons with Ranking.

In a ranking test, the judge is asked to rank the samples in order according to the intensity of some particular flavor factor. For example, when testing the use of dehydrated onion flakes versus concentrated onion juice, the ballot might read, "Rank the samples according to the intensity of onion flavor present. Place the number of the sample having the most intense onion flavor first and the sample with the least onion flavor last". In this type of rank test, the magnitude of difference between samples is scored in equal divisions which actually might not be the case.

4. Multiple Comparison with Scoring.

In scoring, a scale is designed to measure the degree or intensity of the particular flavor factor being tested. A ballot designed to measure the effect of cooking method on the tenderness of meat might read, "Very tender, moderately tender, slightly tender, slightly tough, moderately tough, very tough". Each sample is supposedly scored as an absolute unit according to a pretrained opinion of the judge.

5. Multiple Comparison, Reference-Difference Scoring.

A known reference or standard sample is so labeled and the coded samples scored in direct comparison to this standard or reference sample. A blind or coded reference sample is included with the other coded samples for scoring. Degrees of difference may be evaluated depending upon the wording of the written ballot. A four-point scale reading "Same flavor, slightly, moderately, extremely different flavor", has proven

satisfactory in the flavorium for most difference tests. However, this scale may be extended to include more divisions or points if desired.

For preference testing, the following methods of testing are probably those most commonly used:

1. Paired Comparison.

The judge is presented with two samples and asked to indicate which sample he prefers. This method may also be used for difference testing and in this event the judge would be asked to indicate which of the two samples was sweeter or more pronounced in onion flavor, etc. The paired comparison test, as usually used, does not give a degree of like or dislike for the sample or a degree of difference. In other words, a judge may indicate the sample he prefers and actually dislike both samples. However, this method is readily understood by the untrained preference judge and is an easy test to administer.

2. Multiple Comparison Scoring Scales.

Many different types of scales have been developed and used to try to determine a degree of like or dislike for a food. These scales may be worded "excellent, very good, good, poor", or in some similar manner. However, the preference scale which has probably received the most attention and use in the last five years is the hedonic scale developed by personnel at the Quartermaster Food and Container Institute. Under the direction of Drs. David R. Peryam, Frances J. Pilgrim and N. F. Girardot, the Quartermaster Corps expended much time and effort to determine just what words would best

express a person's like or dislike for a food. The hedonic scale is the result of these investigations. It is a nine-point scale worded as follows:

Like extremely
Like very much
Like moderately
Like slightly
Neither like nor dislike
Dislike slightly
Dislike moderately
Dislike very much
Dislike extremely

Score points from 1 to 9 are assigned to the terms for the purpose of statistical analysis.

3. Multiple Comparisons with Ranking.

This is the same procedure as mentioned under difference testing except when used as a preference test the ballot is worded so the judge will indicate in order which sample he prefers or likes.

4. Multiple Comparison, Reference-Preference Scoring.

This test is similar to the reference-difference test except for the wording of the scale to indicate a like or dislike for the sample. This method gives a direct comparison if a known standard or reference sample is available. A blind reference is also included for scoring with the coded samples. The judges are asked to score the coded samples in direct relation to the reference sample as to whether the flavor is neither better nor poorer than the reference sample, or slightly, moderately, very much or extremely better or poorer than the reference sample. This method of testing is very helpful in explaining the results of the test to people unfamiliar with flavor testing procedures.

No one type of test is best or most suitable for all flavor testing. Careful consideration must be given to the objectives of the study and the type of test which will best answer these objectives. The wording of the printed ballot on which the judges score their opinions must be carefully constructed as it is the key to the meaning of the test.

Any test should be so designed that the results can be statistically analyzed. However, an analysis of the means does not always tell the complete story. A frequency table or table of distribution of scores along with the mean scores is necessary to determine the complete answer. An example of the importance of the frequency table can be shown by the results of a panel conducted in the flavorium on canned green beans containing 0 and 0.15% monosodium glutamate. This was a seven-point reference-preference scale with college students used as judges. There was no significant difference between the mean scores of the samples. However, the frequency table showed that on the MSG-treated sample, the number of judges divided from the center point in this manner:

	0	0.15% MSG
Extremely better flavor	1	2
Moderately better	3	14
Slightly better	16	30
Neither better nor poorer flavor	51	17
Slightly poorer flavor	22	23
Moderately poorer	6	11
Extremely poorer flavor	1	2
Total Judges	140	140
Mean Scores	3.9	4.1

The judges as a group did not show a preference for either sample. However, from the frequency table, it is evident that the individual judges could detect a difference in flavor between the treated and untreated samples but were equally divided as to whether the MSG sample was better or poorer in flavor than the untreated sample thus giving a mean score which indicated no preference.

The results of any flavor test are only as valid as the samples and procedures used in conducting the test. Careful consideration must be given the complex of factors involved when using human subjects as test machines. Only by careful planning and realization of the psychological and environmental factors involved can valid results be obtained. However, there is no reason why any laboratory which is willing to give time and thought to the problem cannot be successful in conducting laboratory flavor panels.

Attached is a selected bibliography on flavor panel studies. This list should be useful for those desiring further information on this subject.

A SELECTED BIBLIOGRAPHY ON FLAVOR TESTING PROCEDURES

Bennett, Grace, Spohr, Barbara M., and Dodds, Mary L.
The Value of Training a Sensory Test Panel.
Food Technology, Vol. 10, No. 4, 205, April, 1956.

Bradley, Ralph A.
Some Statistical Methods in Taste Testing and Quality Control.
Biometrics, Vol. 9, No. 1, March, 1953.

Byer, A. J. and Abrams, D.
A Comparison of the Triangular and Two-Sample Taste-Test Methods.
Food Technology, Vol. 7, No. 4, 185, April, 1953.

Dawson, Elsie H. and Harris, Betsy L.
Sensory Methods for Measuring Differences in Food Quality.
Agricultural Information Bulletin No. 34, U.S.D.A., August, 1951.

Filipello, F.
Organoleptic Wine-Quality Evaluation I. Standards of Quality and Scoring Versus Rating Scales.
Food Technology, Vol. 11, No. 1, 47, Jan., 1957.

Foster, Dean.
Approach to Panel Studies of Foods and Needs for Standardization.
Food Technology, Vol. 8, No. 6, 304, June, 1954.

Girardot, Norman F., Peryam, David R., and Shapiro, Ruth.
Selection of Sensory Testing Panels.
Food Technology, Vol. 6, No. 4, 140, April, 1952.

Gridgeman, N. T.
Taste Comparisons: Two Samples or Three?
Food Technology, Vol. 9, No. 3, 148, March, 1955.

Gridgeman, N. T.
Group Size in Taste Sorting Trials.
Food Research, Vol. 21, No. 5, 534, Sept.-Oct., 1956.

Harries, J. M.
Positional Bias in Sensory Assessments.
Food Technology, Vol. 10, No. 2, 86, Feb., 1956.

Harries, J. M.
Sensory Tests and Consumer Acceptance.
J. of Science of Food and Agriculture, Vol. 4, 477, 1953.

Harrison, S., and Elder, L. W.
Some Application of Statistics to Laboratory Taste Testing.
Food Technology, Vol. 4, No. 11, 434, Nov., 1950.

Mitchell, John W.
The Effect of Assignment of Testing Materials to the Paired and Odd Position in the Duo-Trio Taste Difference Test.
Food Technology, Vol. 10, No. 4, 169, April, 1956.

Mitchell, John W.
Time-Errors in the Paired Comparison Taste Preference Test.
Food Technology, Vol. 10, No. 5, 218, May, 1955.

Miller, P. G., Nair, J. H., and Harriman, A. J.
A Household and a Laboratory Type of Panel for Testing Consumer Preference.
Food Technology, Vol. 9, No. 9, 445, September, 1955.

Peryam, D. R., Pilgrim, F. J., and Peterson, M. S.
Food Acceptance Testing Methodology, A Symposium, Quartermaster Food & Container Institute, 1954.

Peryam, D. R., and Girardot, N. F.
Advanced Taste-test Method.
Food Engineering, 194, 58-61, 1952.

Pettit, L. A.
Informational Bias in Flavor Preference Testing.
Food Technology, Vol. 12, No. 1, 12, Jan., 1958.

Pettit, L. A.
Quantity of Sample, Swallowing, and Rinsing Factors in Flavor Preference Testing of Tomato Juice.
Food Technology, Vol. 12, No. 1, 1, January, 1958.

Pilgrim, Frances J., and Wood, Kenneth R.
Comparative Sensitivity of Rating Scale and Paired Comparison Methods for Measuring Consumer Preferences.
Food Technology, Vol. 9, No. 8, 385, 1955, Aug.

Roessler, E. B., Warren, J. and Guymon, J. F.
Significance in Triangular Taste Tests.
Food Research, Vol. 13, No. 6, 503, Nov.-Dec., 1948.

Studies in Food Science and Technology
I. Methodology of Sensory Testing.
Food Technology, Vol. 11, No. 9, Sept., 1957.