Flavor Evaluation of Fats and Oils

C. D. EVANS, Northern Utilization Research Branch, Agricultural Research Service,
U. S. Department of Agriculture, Peoria, Illinois

"To dispute taste is futile" is an old Roman expression. Until recently the role which taste plays in the selection of foods has been minimized. Taste and smell are the only two human senses which have not been relegated to a secondary position in modern processing operations. The development of automatic mechanical control and sensing equipment has not only displaced the other human senses but has brought about new standards of quality and uniformity. Even though the human eye is reported to be able to distinguish over 7,000,000 colors and shades, it cannot compete with the modern recording spectrophotometers. The physiologist refers to taste and smell as the lower senses, in contrast to the senses of sight and hearing, which are highly developed in man. The functioning of these lower senses are incompletely understood, and we do not have adequate physical laws or measurements by which we can judge them. The actual mechanism of receptor stimulation is not known (53).

C. D. Evans

Flavor in foods is the combined result of the senses of taste and smell, plus those of touch, temperature, and pain. Taste and smell result from contact stimuli, where the stimulating substances must be placed upon the receptive sensory cells. The primary tastes are usually given as sour, salty, bitter, and sweet although some physiologists also include alkaline with this group. The complex tastes of many stimuli are supposedly duplicated by appropriate mixtures of substances that elicit the four basic qualities. The taste-sensing organs are spindle-shaped cells grouped together on the tongue in goblet-like clusters called taste buds. The taste buds are primarily located on the dorsal surface of the tongue and in trenches at the back of the tongue. The middorsal section of the tongue is devoid of taste buds and is insensitive to flavor stimuli. Taste buds are found to a lesser extent on the palate, pharnyx, and larnyx. Children have the greatest number of taste buds and with increasing age the numbers decrease. The reduced sensitivity found in older people may be associated with this decrease in the number of taste buds.

The sense of odor is much less understood. Little fundamental knowledge is available because of the difficulty in conducting experimental work with the comparatively inaccessible odor receptors. The olfactory receptors are located in two small patches of yellowish-brown pigmented membranes at the top of each nasal cavity. These receptive cells, grouped in an area of about one square inch, form the nerve fibers which lead directly to the olfactory bulb at the base of the brain. During breathing, odorous particles are carried through the nasal passages, and eddy currents cause the air to mix and be swept over the nerve fibers. The expiration of air also carries odors to the nerve cells in a similar manner. Much of the odor stimulation during eating takes place in this latter manner. In terms of concentration the sensitivity of olfactory organs is some 10,000 times greater than the taste organs (43).

Ethyl alcohol has acquired somewhat of a reputation as a stimulant. It also is one of the few substances which has odor stimulation and a taste stimulation independent of odor. Moncrieff gives the following concentrations as the lowest that can be smelled and tasted (Table I).

TABLE I
Odor and Taste Thresholds

Ethyl alcohol
0.44% wt./wt. in air for smell
14.00% wt./wt. in water for taste
Ethyl mercaptan
3×10^{-9}% wt./wt. in air for smell
Strychnine hydrochloride
4×10^{-5}% wt./wt. in water for taste

Ethyl alcohol is comparatively weak in odor and taste, but ethyl mercaptan is extremely odoriferous. It can be detected at concentrations 100 million times lower than alcohol. Strychnine hydrochloride is one of the most bitter substances known, and it can be detected at concentrations of 4×10^{-5}%. Short-chain aldehydes and fatty acids are some of the breakdown products of fat oxidation, and these can also be detected at these extreme dilutions. The Northern Utilization Research Branch panel detected butyraldehyde dissolved in mineral oil at a concentration of 6.6×10^{-6}%. Few people exercise their sense of smell to anywhere near the capacity of which they are capable. In taste-panel evaluations every effort should be made to enlarge the use of this latent and potent power. The remarkable analytical ability of the trained perfume chemist to distinguish over 1,000 odor components attests to the sensitivity of smell (50).

Sensory Panels

Three or four types of taste panels are generally recognized and their purposes are primarily different. This discussion will be confined to the problems of the analytical-type panel concerned primarily with the detection of a difference between samples.

Quality control panels are usually quite small and efficient and are used for the maintenance of standards of production. The establishment of official grades is usually done by a small number of graders with long experience in the field of their particular product. Consumer preference panels are large untrained groups from which preference decisions are obtained without the use of any standards. Morse (44) has recently reviewed consumer preference studies and market research surveys.

Several very extensive bibliographies have appeared since 1950 on taste panels and sensory testing (1, 6, 19) and on statistical methods for sensory difference testing (9, 10, 11, 12).

The need for discriminating types of taste panels usually arises from problems on quality improvement of a product or on improvements in processing techniques. Wide use is also made of analytical panels in solving the numerous problems which arise in the development of any new type of product. Many panels have been established in the fat and oil industries to evaluate salad oils, shortenings, margarine, chocolate liquor, peanut butter, and similar materials. These are usually highly trained panels being used as analytical tools to evaluate small differences which ultimately lead to the production of improved products. Quantitative taste-panel results are of equal importance to research, production, and sales departments.

Quality improvements in flavor, stability, and appearance are the usual factors considered by taste panels. The effect of processing variables on quality can be quantitatively evaluated by adequately designed experiments with taste panels. Equally valuable results are obtained when a reduction in processing time can be shown or when material savings are accomplished in raw materials used in the manufacturing process. One edible oil processor has reported an 80% reduction in time of deodorization without harmful effects on either the quality or stability of the oil (18).

At the Northern Branch we have made extensive use of the analytical type of taste panel in studies designed to improve the flavor stability of soybean oil (22, 45). These investigations have made use of taste-panel data to show the effect of each processing step on the quality and stability of the oil. Similar studies to attain optimum processing conditions have been made by many industrial plants, and we have cooperated with them in helping to establish their taste panels. Taste-panel results have emphasized the importance of, and the critical control which must be exercised over deodorization if a quality oil is to be produced. The effect of trace metals, light, and storage conditions on oil quality and stability has been extensively studied in our laboratory. Trace-metal contamination is one of the most detrimental factors contributing to poor quality. Iron can be added to soybean oil and off-flavors detected in the freshly deodorized oil at approximately 1 part in 10 million. To taste an off-flavor in your drinking water through the addition of iron chloride requires that you add about 1,000 times as much, or a concentration of about 100 p.p.m. Such metals as copper and cobalt are even worse offenders in destroying the quality of fats and oils.

Many taste-panel studies are concerned with quality improvement of fats through the use of stabilizers. These additives may be antioxidants, metal inactivators, phosphatides, or synergistic mixtures and may in themselves impart a flavor to the oil. We have found, for example, that phosphoric acid cannot be added to oils above a concentration of 20 p.p.m. without atypical flavors developing. The amount of lecithin that can be added is limited by the development of detrimental color and taste. The concentration detectable by taste is roughly equivalent in terms of phosphoric acid to that given above.

Although most of the literature on flavor evaluation of oils is concerned with liquid or salad oils, we have found that the technique is equally applicable for evaluating margarine base stocks and shortenings. These oils are served to the panel as liquids and are handled and evaluated just as are the salad oils. Obviously, only fats with melting points below body temperature can be served in this manner. Some of the mono- and diglycerides which melt slightly above body temperature will congeal rapidly in the mouth to give an unpleasant tallowy sensation.

Sample Presentation

In the paired-sample technique the judges are presented with two samples which they are asked to score. The design of the experiment is usually simple but such that every sample of the experiment is compared with every other sample. This type of comparison, although direct and easily understood and interpreted, is time-consuming. The time expended in conducting the tests and the low probability of obtaining results of significant differences are the main objections. Its simplicity is an attribute, and the effects of many unknown influences are made obvious in such comparisons. It is well known that a sample is scored much higher when compared with a poor sample and scored lower when compared with a better sample. Taster inconsistencies and performance are probably more easily evaluated and observed in these tests because the taster is required to render fewer judgments and has less cause for guessing.

Byer and Abrams (14) found that in taste tests their panel showed a more significant discrimination in the paired-sample test than in the triangle test. Pfaffmann (54) confirmed the results of Byer and Abrams in that the 2-sample test is superior to the triangle method when the flavor dimension can be specified. When the flavor dimensions by which the samples differed were not specified, the more complex triangle test was not inferior to the paired method. Other studies by Gridgeman (26) in the comparison of 2- *vs.* 3-sample tests showed that the 3-sample tests were not normally superior to the paired tests. The probability of correct discrimination was distinctly higher for the paired tests over both the duo-trio and triangle test. In addition, these studies also showed that the paired tests were not as inefficient of time as many panel operators have assumed.

The duo-trio test of discrimination was developed by Peryam and Schwartz (51) and involves the presentation of three samples simultaneously. The judge tastes each sample in left-to-right order and then decides whether the second or third was like the first. Since the odd sample might occupy either the second or third position, the probability is 50% or one-half. Thus four combinations are possible—ABA, AAB, BAB, and BBA—and these are presented in random order, equally often, to complete the test. Like the triangle test this situation calls for discrimination only and reportedly gave excellent results in the investigation of dried milk and in the selection of taste-panel members of superior sensitivity. Other workers (26, 54) have not found the test superior to either the paired test or the triangle test.

The development of the triangle test is usually attributed to Helm of the Carlsberg Breweries, Copenhagen, Denmark, where it was used for control work and for the selection of taste panels. Because of the higher probability of success the test has appealed to many through the consequent saving in time and the number of samples which need to be presented to the tasters. Three samples are presented to the taster; any two are identical. The taster is told that it is a triangle test, and it is his problem to indicate the odd sample. Many times the taster is also asked to indicate the distinguishing characteristics by which the odd sample was identified. This, of course, in essence amounts to scoring of the sample, and the triangle test if correct, serves to substantiate the reliability of the taster. The taster's problem in the triangle test is solely to pick the odd sample; under such conditions he is at liberty to use any and all his facilities in picking the odd sample. The control of the test must therefore be carefully handled so that differences in color, appearance, sample size, etc., will not enable the judge to pick the odd sample without even tasting or smelling. The probability of success by guessing is one-third since one sample out of three must be selected. However there are six ways of presenting the samples, thus ABB, AAB, and ABA, and the reverse BAA, etc. In applying the test all combinations should be presented and in a random order, using each combination approximately the same number of times.

In ranking tests the judges are asked to arrange a series of samples in a decreasing or increasing order of some characteristic. Ranking avoids the difficulties usually involved in the selection of a suitable scoring scale and the selection of adequate standards. The usual criticism of ranking is that in quantitative evaluations the magnitude of the difference in samples is lost. It is our experience however that when differences between samples is small, tasters may not score them differently, but they can easily arrange them in a rank order. With small differences, ranking procedures are much simpler than attempts to revise the scoring system and scale. Ranking also has the advantage of reducing the effects of "erratic" tasters who have strong dislikes or likes for certain flavors. Thus in scoring such an individual might give a zero score to a sample which the rest of the panel regards as average. In ranking, any sample can be reduced only to the bottom position, which gives a much smaller reduction to the panel average. Ranking is also helpful in keeping in line those tasters who always try to "beat the game" by scoring samples the same when small differences exist and scoring extreme differences when the panel average indicates that only a normal difference exists. Some tasters in "beating the game" will not use the full scale but limit their scores to within the known range of the panel average. Tasters of the last two categories should be eliminated from difference testing panels.

Ranking is also preferred when a preference judgment is required from a large untrained panel. Ranking techniques have been extensively employed in psychology and all types of food testing (8, 19).

Handschumaker (27) designed a ranking test for the study of flavor reversion in soybean oil shortening. In this series there were five controls and the unknown sample. Tasters were required to rank the series by odor alone. It was found that those unable to rank by odor were not helped by tasting of the samples. The lack of improvement with tasting was attributed to low sensitivity of the individuals or to the loss of acuity because of the large number of samples to be tasted. The use of five controls allowed a check on each taster. The results of tasters who could not arrange the controls in the proper order were not considered in the final score. The use of a large number of controls has the advantage where only a limited number of personnel are available as panel members. In a small laboratory a large number of tasters on the panel are likely to be involved in the program or are at least familiar with it, and such a series of controls will limit personal prejudice. Others have suggested more than one control in a series (16, 29). Terry *et al.* (61) have developed a rank-order method which considers the comparison of two or more products or treatments. The test is flexible, and statistical calculations are made from the summation of ranks. In the simple test of three samples every taster ranks by pair, each sample against the others. Thus for three samples three-paired rankings are feasible. The rankings of each taster are summed (two rankings of each sample), then the rankings of all the tasters are summed. These sums are arranged in ascending order and from the tables significant differences are obtained. Tables are constructed to contain all possible ranking combinations, and the authors expect to extend the tables to include designs for ranking three treatments within incomplete block designs.

In simple ranking with the highest rated sample ranked 1, the next highest 2, etc., the sums of ranks will vary inversely to the numerical scores. The ranked sums follow a normal probability distribution, and for 10 or more tastings the usual statistical difference tests such as the "t" test are applied.

Two samples may be presented to a taste panel for difference testing in any number of ways, and probabilities much greater than one-half or one-third can be obtained. These multiple variety tests are discussed by Wood (63), Lockhart (39), and Evans *et al.* (24); however no application of such tests has been made in the evaluation of fats because of the large number of samples that must be presented to a taster at one time. The selection of members for a coffee-tasting panel, using a multiple-selection test, is reported by Harrison and Elder (29).

Methodology of Testing

The satisfactory operation of an analytical taste panel requires adequate physical facilities. Quiet and pleasant surroundings are absolutely essential for the taster to concentrate on the problems of odor and taste. Individual panel booths where the taster is free from interruptions, suggestions, comments, and facial expressions of other tasters adds to the reliability of the results and to the smoothness of operation of the panel. Noisy disturbances markedly detract from the abilities of tasters who are normally not bothered by them. Such occurrences as whistling, banging of a door, sliding of a chair, etc., will upset the delicate sensing reactions of a concentrating taster. Many food industries with extensive taste-testing programs have established separate air-conditioned rooms for conducting sensory tests under the optimum of conditions. The temperature, humidity, and light conditions are all under control in these rooms. To eliminate small differences in color, brilliance, or sheen of products under test many taste-panel booths are constructed so that the intensity of the light and its wavelength are under control of the taste-panel operator. Attention to many fine details and an interest in people and their reactions are part of the requirements of a good taste-panel operator.

Panel Selection

Since the reliability of a taste panel's results is determined by statistical analysis, panels should have at least 10 members. To attain statistical significance repeat testings will be required of panels having fewer members. Any method of selecting tasters should include a preliminary training period to acquaint the tasters with the quality factors involved. The tasters should be trained on the type of material they will evaluate. Following the training period, a period of blind testing is conducted to establish the reliability and discrimination of the individual tasters. The first panel established at the Northern laboratory was selected on the basis of acuity tests for the four basic tastes, following a procedure for establishing food panels (47). Sixty per cent of the individuals were eliminated because of high thresholds or incorrect identifications of the basic tastes. After some experience in evaluating oils the final selection was reduced to 23% of the original group surveyed for tasters. At a later date when replacements to the panel were required, training was limited to three weeks in which only oils were evaluated (45). As a result of this training two-thirds of the people were found to be satisfactory panel members. It is believed that high acuity for the four primary tastes is not essential in order that a person may be a good oil taster. Mackay and Jones (41) found that high acuity did not correlate significantly with the tasters' ability to rank foods containing varying amounts of the basic taste stimulants. The low percentage of tasters obtained in our first survey probably results from the low acuity of several of the prospective tasters and from our lack of experience in conducting taste-panel operations. Hanson (28) has observed that about 30 to 40% of technical personnel make good taste-panel members. Page and Lubatti (49), in selecting members of a panel to determine when food flavors were tainted by fumigants, stated that only 12% of the individuals show outstanding ability. Schlosberg *et al.* (56) in a study of taste-panel selection and training obtained no clear-cut evidence that selection or training had any effect on the performance of either a difference testing or a preference testing panel. These results were obtained on an undergraduate college population quite homogenous in regards to background intelligence, age, diet, and motivation.

Sample Size

Since taste-panel operations involve the expenditure of considerable time and effort of the personnel connected with the tasting program, the greatest efficiency can be obtained by tasting the maximum number of samples at each session. Burrows (13) felt he could evaluate 10 samples of good fat at one session. MacLean and Wickens (40) report a loss of taste perception in the evaluation of several samples of cocoa-bean liquor. Studies with the Northern laboratory's panel showed that 6 samples of oil were too many. Other tests showed that the panel was just as sensitive when scoring 4 samples at a time as when only 2 were presented. The panel members however were very much against scoring more than 2 samples per session as they felt that they had lost considerable of their perception before tasting the fourth sample. Four samples are the maximum presented, and they should all be of good quality. The rate of deterioration in taster performance occurring during the tasting of a large number of samples depends on the type of food under test. It is well known that the taste and odor senses will adapt to certain flavors much more rapidly than others. Results with coffee indicate no tendency toward fatigue while tests with maple sirup showed a significant decline (29). Pfaffmann reports no loss in discrimination with several foods even after presenting 50 to 75 samples at one session (54). Sulfited foods have been observed to give erratic results because of the dulled acuity of the taster. Boggs and Ward (7) reported that the tasting of one sample of potates containing 12 to 100 p.p.m. of sulfur dioxide dulled the perception for a second sample containing sulfite. Neubert and Carter (48) found that the tasting of a single sample was superior for the detection of foreign flavors found in apple juice which had been prepared from demeton-sprayed apples. Indiscriminate use of chemical defoliates, herbicides, and fumigants has produced many a flavor problem.

We have found that SO_2 is detectable in liquid soybean oil at a concentration less than 3 p.p.m., which is considerably lower than that discernible in other foods. Odor scores would indicate that it was not detected by smell at this level. Sulfur dioxide has been proposed as a fractionating solvent for soybean oil. Very specific and interesting taste problems could develop with such a process.

Thus the material under investigation has a lot to do with taster fatigue and adaptation, and the number of samples that can be satisfactorily evaluated. Odor in oils, like other foods, can be detected more readily in warm oil than in cold oil. Also the mouth feel is much more pleasant when the oil is warm, and it does not detract from the taster's concentration on odor and flavor. The usual procedure in tasting oils is to score on odor first and then taste the sample having the preferable or least odor. To develop odor the samples of oils are heated and served to the tasters at 45°C. Adequate sample size of not less than 7.5 to 10 ml. should be served to each taster. An amount large enough to cover all the mouth surfaces should be taken into the mouth and mulled about. Tasters are always instructed not to swallow the sample regardless of the quality, and knowing that it is not to be swallowed they will take a larger amount of oil into their mouth. About 7.5 ml. are adequate for a single tasting; repeat tastings of the same sample contribute nothing but taster fatigue.

Warm water is recommended for rinsing the mouth between samples although slices of bread, unsalted crackers, and apple slices have also been used. After rinsing sufficient time should be allowed for the saliva again to bathe the taste buds.

Score Sheets

An appropriate score sheet should help the taster in making his evaluations. The simpler the score sheet, the better. As a generalization it can be stated that the fewer the judgments required of a taster the less confusion, and the more reliable are the results. In taste-panel operation a written answer should always be required. Score sheets vary with each problem and the material under investigation. The simplest scoring systems require a checking of yes or no answers, or a ranking of a short series while the highly complicated systems score a multiple of factors including flavor and off-flavor descriptions. Many taste-panel operators find it more satisfactory to use a descriptive score sheet and to eliminate any attempt at scaling or ranking by the judges (15, 52). This is especially true when fixed standards are not available. The score sheet consists of a long list of adjectives describing the attributes of the material under study. A judge checks the adjective or adjectives which he believes best describes the material; he may also indicate the degree of intensity, *i.e.*, weak, moderate, or strong. Later the taste-panel operator assigns a numerical score based on the judges' description of the material. The numerical scores are handled in the usual statistical manner for comparison of significant differences.

In evaluation of purified fats, except for olive and perhaps corn oil, the highest standards are for the most colorless, odorless, and tasteless product obtainable. The "ideal" material and thus the best scoring material should be easily recognized. However should a corn oil processor desire a certain level of flavor typical to this particular oil, then the ideal scoring material can be recognized only after an extensive training period coupled with considerable taster experience. Thus the scoring system which Hopkins (32) has designed would be very applicable for evaluation of corn oil. This so-called "balanced scoring system" is based on intensity scores of attributes from -5 through 0 to a $+5$. These 11 graduations vary from a gross deficiency (-5) to a gross excess ($+5$) with 0 being the normal, or the ideal level. The use of this test has been applied to the evaluation of oils by Lips (38). Although the system is applicable to a wide variety of food products, only half of the scale will be used when the presence of any odor or flavor is detrimental to the quality score of the product.

The number of units in the scoring scale should depend on the ability of the individual tasters to detect differences between each successive unit. Most scoring systems will have in the neighborhood of 10 units. Such a number seems adequate, convenient, and generally satisfactory and is most helpful for the many tasters who subconsciously score in terms of 100. Dr. Dove explains this 100-unit idea as a grading system carried over from grammar-school days; it may also result from common thinking in terms of percentage. Uniformity in scoring is essential in analyzing the results, and usually a decrease in number of grades improves the uniformity of results. However the scale cannot become too coarse, or it will lose its discriminating value. A 10-unit score allows the expression of quality in subjective terms so that there is no contradiction in their meaning. The adjectives should be simple, concise, and understandable terms which retain the desired proportionality between score and quality. Difficulties often arise in assessing total food quality by combining scores made for specific characteristics. Various ways of combining such scores have been suggested, such as the product, geometric mean, or harmonic mean in place of simple summation of scores (23, 33, 63).

Interpretation of Results

Statistics are primarily a measurement of error or, as Dr. Snedecor states it, "the art of evaluating the uncertainty of your judgments." Statistical methods are a tool to explain and interpret the data in the light of the variation (errors) and probabilities involved. Statistical analysis cannot increase the validity of the data. Sound methods of analysis are covered in the several elementary texts on statistics, and application by example is given in almost all the papers dealing with organoleptic analysis. The number of correct answers to establish significance in the triangle test was tabulated by Helm and Trolle (30). Application of the Chi square test is given by Boggs and Hanson (6), and the use of the critical ratio to establish degree of significance is discussed by Peryam and Swartz (51). Harrison and Elder (29) have published charts of levels of significance for both the paired comparison and triangle test covering a wide range of probabilities and panel sizes. Moser *et al.* (47) have discussed the application of the "t" test to paired samples. Terry *et al.* (61) and Bradley (8) have discussed a number of rank-order tests and their application. The transformation of ranks to scores is discussed by Bliss *et al.* (4).

There are many experimental designs available with various methods of statistical analysis applicable to the numerical data. In paired comparisons the interpretation is based on the mean score of the panel which is calculated for each sample along with the error term. The error term is usually expressed as the variance of the mean or standard error, and it is used in calculating the probabilities that the score of sample A differs from sample B by a difference greater than can be obtained by chance. The term "significant" result means that the difference between scores is greater than can be expected at a rate of 1 out of 20 trials. Stated somewhat differently, a "significant" result means that if you will accept the assumption that the observed difference in scores is real, you will be wrong only once in 20 times. The occurrence of an event once in 20 trials is a probability of 0.05 or 5%. Significant probabilities are usually indicated by a single asterisk (*) and mean that the significance of the results lies somewhere between the 5% and the 1% levels. Highly significant results, *i.e.*, above 1%, are indicated by two asterisks (**).

The literature survey of Dawson and Harris (19) indicates that the most common method employed by experimenters in analyzing sensory data has been the analysis of variance and correlation. Bradley and Somerville (10) have stated that the ideal scoring scale should have five attributes if standard methods of statistical analysis can be applied. Bradley (8) and Terry *et al.* (61) also discuss the validity of using analysis of variance techniques. The basic assumptions for an ideal scoring scale are a) observations are continuous, b) random fluctuations are normally distributed, c) observations are independently distributed, d) error variances are homogeneous, and e) treatment and environmental effects are additive.

Motivation

The attitude of the tasters is the one factor upon which depends the successful operation of any taste panel. A dissatisfied or disgruntled person is no longer a reliable taster. Although he will go through the formalities of tasting, results are only half-hearted and often of questionable reliability. Bengtsson and Helm (3) state that interest is the prime attribute of a good taster. Boggs and Hanson (6) discuss the problem of creation and maintenance of interest and the avoidance of prejudice which may result where too much knowledge is imparted to the tasters about the problem or samples under investigation. When a distinct level of difference is to be determined by an analytical panel, considerably more information can be given the tasters than can be imparted when levels of acceptability or preference are demanded of the panel.

Techniques for increasing interest are recognized and employed by all investigators. These may take the form of panel discussions, where the problem is fully presented and methods of solution and the details of the taste test are outlined. The giving of monetary rewards has frequently been employed. Other rewards such as a coffee break, serving of cookies, etc., have proven helpful. We have found that a small bulletin board containing clippings about or of interest to panel members, graphs of current commodity prices, even jokes and cartoons, add their beneficial psychological effect.

Identification of the samples and the scores of tasters are made available to the panel members immediately after tasting. This comparison and discussion of the results among panel members and with the panel operator has helped to improve performance and maintain interest in the work.

Recently Pfaffmann (54) has conducted a series of experiments which showed the beneficial effect obtained in panel discrimination when an immediate knowledge on the correctness of an answer was given to the taster. The "knowledge" group improved in the number of correct judgments rendered to 84% while the control group gave only 65% correct answers. Previous scores of the two panels on the same substance had been 69 and 71%, respectively. To check this performance the test was repeated with an inexperienced group with essentially the same results. The author concluded that the panel members given immediate knowledge of their results substantially improved their performance during the course of the experiment. Improved performance was demonstrated with experienced or beginning panels and with either difficult or easy discriminations. Improvement was attributed to motivation rather than the ease of learning. Although motivation as such was not listed by Bengtsson and Helm (3) as one of the 10 rules of taste testing, I believe it to be very essential for continued and successful taste-panel operation.

Supplementary Chemical Tests

Chemical and physical tests are more reproducible and less time-consuming than are the sensory tests; because of this there is a constant desire to eliminate or minimize the amount of taste-panel work. Beadle (2) in his discussion on fat stability points out that, in the final analysis, rancidity must be determined by organoleptic observation. Many correlations have been made between organoleptic scores and various chemical tests. At the present time the peroxide determination is the most widely used chemical test for the determination of fat quality. Peroxide values have shown a good correlation with the organoleptic flavor scores of soybean oil (21). Similar studies on cottonseed and safflower oil have been carried out in our laboratory, and essentially the same degree of correlation was obtained between the flavor score and peroxide values. Grant and Lips (25) studying rancidity in lard correlated odor scores with seven different chemical tests and found that the association of organoleptic scores was best with peroxide and alpha-dicarbonyl values. It however has been pointed out many times that a peroxide determination does not give a full and unqualified evaluation of fat quality. After a limited amount of experience is gained with a particular sample of oil, then a fair prediction of the flavor score can be made from the peroxide values. High peroxide scores usually mean poor flavor scores, but a low peroxide value is not necessarily an indication of a high flavor score. We have found it possible to oxidize soybean oil rapidly under oxygen at 60°C., and if the oil is tasted immediately, peroxide values may go as high as 10 before a marked drop in the flavor score occurs. Soybean oil that is allowed to oxidize under normal conditions will have an extremely low flavor and odor score when a peroxide value of 10 is attained.

Aldehydes have long been identified as oxidation and breakdown products of fats. Although the Kreis test is no longer widely used in fat evaluation work, the isolation and identification of several specific aldehydes as oxidative products of fats has again given impetus to carbonyl tests (35, 42, 55, 60). A quantitative carbonyl method (37) capable of determining concentrations of 5×10^{-6} molar (equal to taste sensitivity) has been employed by several investigators in fat oxidation studies. Henick *et al.* (31) have modified and adapted the method for use on fats and fatty foods and included in the determination both saturated and allenic carbonyls. Sidwell *et al.* (57) has compared total aldehyde content, peroxide value, and thiobarbituric acid values obtained from the oxidation of several fats. The authors included in their discussion the relationship of these values to the sensory values obtained with a trained taste panel. Chang and Kummerow (17) have used the carbonyl determination as a basis for developing an instrumental method for measuring the degree of reversion and rancidity of edible oils. Carbonyl indices were found to correlate very well with the organoleptic scores of various aged liquid oils and aged hydrogenated oils.

Thiobarbituric acid test has found extensive use in the past few years as a measure of fat oxidation (5, 20, 57, 58, 62). Although the chemistry of the reaction is not known (34), these reports indicate a high degree of correlation of TBA values with organoleptic scores. Biggs and Bryant (5) report studies on butter, whole milk powder, and cheese and state that the test is capable of measuring the degree of oxidation below the level of organoleptic sensitivity. Dunkley (20) showed that the TBA test correlates closely with numerical flavor scores of milk samples having oxidized flavors of varied intensity. Turner *et al.* (62) report the method to give a more reliable index on rancidity in frozen pork than any other chemical test and report a significant correlation between taste acceptability scores of wieners and pork patties with the TBA value of the pork used. Sidwell *et al.* (57, 58) has shown a direct relationship between flavor scores and TBA values for fats and dried milk products. Kenaston *et al.* (36) has shown that the TBA test is the most sensitive chemical test for fat oxidation products. It is 30–80 times more sensitive toward linolenate oxidation products than linoleate, and it is practically negative toward oleate oxidation products. Since the unsaturated acids in cocoa butter are practically all oleic acid, a correlation study of the flavor scores and TBA values with oxidation of this fat would be of considerable interest. The inference might be gained from a review of the chemical stability tests that all oils have a very high initial flavor score. This is not true, as is well known by any oil processor. Figure 1 shows the variation in opinion

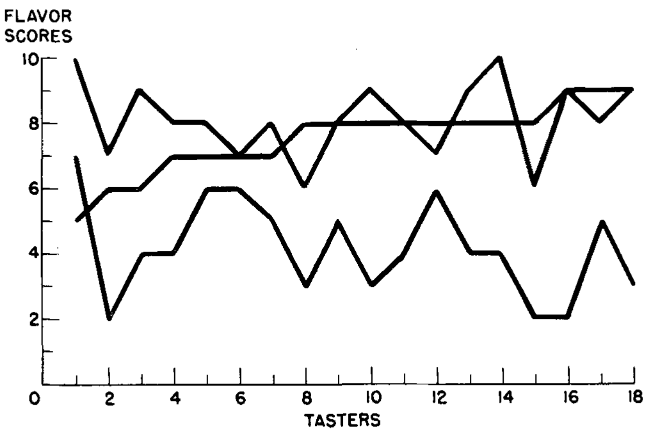

FIG. 1. Taster variation shown in the evaluation of three soybean oils.

of 18 tasters on the flavor scores of three freshly deodorized soybean oils. These three oils would react negatively to any of the chemical stability tests and therefore would all be rated high and equal. There are many uses for an organoleptic taste panel besides studying the rate or the extent of flavor deterioration of an oxidized fat.

Taster Performance

The variation in the sensory acuity of the individual trained taster is a constant source of error. The extreme variation which is shown repeatedly in taste-panel evaluations may be very disconcerting to the panel operator and to the executives of the organization. The individual errors however tend to be compensating in panels of adequate size, and the over-all error of the panel can only be reduced to a minimum through taster selection and training. The individual variation which may be expected in the tasting of oils is shown in Table II. Three oil samples were evaluated by the Regional laboratory panel of 13 members, and the same samples were evaluated by 18 industrial chemists who were attending a conference on taste-panel methodology at our laboratory. These visitors, representing 13 companies, were all thoroughly familiar with oil tasting, and many were in control of margarine production for their respective companies. This group would be as representative and as critical of oil quality as could be expected from any polling of the margarine oil industry. The industrial people were familiar with the scoring system and the methodology of testing. Many had served on analytical and quality control panels. The results show excellent agreement between the two panels. Identical scores cannot be expected from two panels composed of so few members. However the results of both panels indicated the same order of quality for these oils. The errors of individual tasters and the range of scores are quite comparable in each panel. A statistical analysis of the results show that each panel found no significant difference between samples 1 and 2, but a highly significant difference was found between sample 3 and samples 1 and 2. The respective scores given by each panel were not significantly different for any of the samples. Such agreement lends confidence to the results and indicates that it would be possible for panels throughout the industry to evaluate and score oils with very satisfactory agreement between panels.

TABLE II
Panel Variation in Judging Edible Oils

Sample no.	Visitors' panel score				NU panel score			
	Panel av.	Std. dev.	Range High	Range Low	Panel av.	Std. dev.	Range High	Range Low
1	8.2	1.2	10	6	8.6	0.9	10	7
2	7.5	1.0	9	5	8.1	1.0	10	7
3	4.1	1.6	7	2	5.2	1.5	8	4

Significance of Results

Sample nos.	Between samples within panel: Visitor	Between samples within panel: NU	Sample nos.	Between panels within samples
1–2	+	+	1	+
1–3	**	**	2	+
2–3	**	**	3	+

In Figure 1 are plotted the individual scores of the visitors' panel for the three oils just considered. If these results were presented to people unfamiliar to taste-panel operation, the usual reaction would be to discard any method showing such a variation in results. These are typical results, with the usual taster error found in day-to-day panel operation. Better agreement in tasters' scores is found with the high quality oils. This is indicated in Table II where the Northern Utilization Research Branch panel's standard deviation drops from 1.5 for a poor oil to 0.9 for a good quality oil and also in Table III which shows how the standard deviation of the Northern Branch's panel varies with the quality of the oil. Data in Table III were obtained from 50 samples in each quality group and indicate that the variation among tasters is almost twice as high in the bad samples as it is in the good samples. As has been stated many times, the degree of goodness is much more easily defined than degrees of badness.

TABLE III
Variation in Standard Deviation with Quality Score of the Oil

Samples	Score range	Average std. dev.
Good	8.0 to 10.0	0.78
Fair	6.0 to 7.9	1.22
Poor	4.0 to 5.9	1.26
Bad	2.0 to 3.9	1.36

Moser *et al.* (46) have recently completed a study on the performance of several tasters over an 8-year period. The data in Tables IV and V are taken from

TABLE IV
Mean Score of Each Taster and Deviation of Individual Score from Panel Average

Year	n	Taster 1 Mean	Taster 1 Dev.	Taster 2 Mean	Taster 2 Dev.	Taster 3 Mean	Taster 3 Dev.
1946	164	6.0	+0.1	6.2	+0.3	5.9	0.0
1947	185	6.7	+0.7	6.3	+0.3	5.7	−0.4
1948	186	6.8	+0.7	6.7	+0.5	5.7	−0.6
1949	167	6.0	0.0	6.1	+0.1	5.3	−0.5
1950	134	6.6	+0.1	6.2	−0.3	6.5	0.0
1951	125	5.4	−0.4	5.7	−0.1	4.5	−1.3
1952	26	5.7	0.0	5.4	−0.3	4.5	−1.2
1953	64	6.6	+0.6	4.8	−1.3	6.5	+0.5

TABLE V

Correlation of Taster's Average Score with Average of Remainder of Panel

Year	n	Taster 1			Taster 2			Taster 3		
		r	b	σyx	r	b	σyx	r	b	σyx
1946	164	0.77	1.18	1.47	0.75	0.80	1.10	0.80	1.21	1.36
1947	185	.82	1.15	1.35	.82	.98	1.18	.84	1.24	1.34
1948	186	.78	1.01	1.35	.76	.88	1.38	.83	1.19	1.32
1949	167	.83	1.24	1.21	.73	.91	1.27	.77	0.84	1.31
1950	134	.77	0.89	1.16	.78	1.01	1.23	.86	1.04	0.96
1951	125	.89	1.15	1.14	.84	1.00	1.26	.85	1.02	1.23
1952	26	.90	1.35	1.06	.80	0.74	0.93	.79	1.12	1.42
1953	64	.66	0.83	1.27	.33	.34	1.33	.78	1.36	1.43

r = Correlation coefficient.
b = Regression coefficient.
σyx = Standard error of regression.

this report, and they show the stability of the individual's average scores as judged by the the panel average. In any single test however the error of an individual may be as high as that shown in Figure 1. The samples evaluated were all soybean oil, which varied in quality from year to year, but every sample was tasted by all the tasters. The average score will vary from year to year, depending on the respective number of fresh and aged samples evaluated. The results show that no taster graded consistently above or below the panel mean over the 8-year period. The maximum deviation was a low of 1.3 shown by two different tasters and a high of 0.7 of a unit shown in two different years by the same taster. To measure the taster's ability to distinguish between oils, the correlation and regression method was used. The correlation coefficient measures the association of the taster scores with those of the panel. The closer the coefficient is to one, the better the relationship. In oil-tasting work we believe that a correlation of 0.85 to 1.00 is excellent, 0.70 to 0.84 good, 0.60 to 0.69 fair, and any value below this is considered poor. These are arbitrary levels and were arrived at through training and taster performance studies. Except for the one exception, Taster No. 2 in 1953, good correlations were obtained. When tested for significance (59) all correlation coefficients were significant at the 1% level except the low figure for Taster No. 2. His correlation coefficient of 0.33 is just significant at the 5% level. The work of Taster No. 2 obviously needs watching, and a correction of his difficulties should be attempted since in previous years he has shown himself to be a good taster.

The regression coefficient indicates the amount of change of the individual score with a unit change in the panel score. The closer this coefficient is to one, the better is the relationship, and normally this relationship is high. Again Taster No. 2 shows the largest variation from the normal behavior. Correlation and regression are related, although the correlation coefficient is most often used in judging the reliability of a taster, actually the regression coefficient is more useful.

The standard error of regression is the error term and shows the amount of variation in the taster scores about the regression line. It can be interpreted like the standard deviation and is a measure of how close the points cluster about the regression line. Thus the smaller this value becomes, correspondingly more faith and reliance can be placed upon the taster's results. The high variation of 1.47 and the low of 0.86 were both obtained by different tasters in the 8-year period over which these data were accumulated. The data on taster performance are presented to show what might be expected of individuals trained as tasters for the evaluation of flavor quality of edible oils and fats. It is of value to those interested in establishing new panels or for the training of new tasters and also for evaluating the performance of existing panel members.

A cooperative oil evaluation study was conducted by our laboratory panel in cooperation with 11 industrial oil taste panels. At monthly intervals two samples of oil were shipped frozen in dry ice to each of the cooperating panels. Twelve sets were sent out during the course of this study, and all were evaluated by using the technique and scoring system of the Northern laboratory. Most panels had approximately 10 members, but there was considerable variation in size both between and within panels, and it was one of the uncontrollable factors in the cooperative program.

Table VI shows the reproducibility of the 12 panels on the quality scoring of a single sample of oil presented 5 different times during the testing program. The standard deviation of the panels varies from 0.14 to 0.97 and for such a small number of panels shows a normal distribution throughout this range. The range of scores of the individuals within all panels was just as high as that discussed previously. The range of scores given by the respective panels is much higher than would be anticipated, on the basis of the known ability of the Northern laboratory's panel to reproduce its evaluations. In contrast to many findings the pooled scores of all the tasters (approximately 130 for each evaluation) showed that the same scores were obtained whether the sample was compared to an oil of equal quality or to an oil of poorer quality. A storage period of 9 months at 0°C. occurred between the conducting of the first test and the last test with this lot of oil. There was no indication of any decrease in the quality score of oil during this storage period. We have obtained other data with our own panel which show that soy-

TABLE VI

Reproducibility of Panels
(Same oil scored 5 times)

Panel No.	Average score	Range of scores	Standard deviation
1	8.1	7.9–8.3	0.14
2	8.2	7.8–8.5	.26
3	8.6	8.2–8.9	.28
4	8.1	7.7–8.6	.35
5	8.1	7.5–8.5	.41
6	8.1	7.7–8.9	.46
7	8.4	7.7–9.0	.48
8	7.6	6.8–8.2	.55
9	7.4	6.8–8.1	.55
10	7.7	6.6–8.3	.76
11	7.3	6.0–8.0	.77
12	6.5	5.9–8.2	.97

bean oil can be stored for about 2 years in full bottles at 0°C. without loss of flavor quality.

Extreme variation was shown in ability of the various cooperative panels to detect signficant differences by the paired sample technique. Some of the data collected is summarized in part in Table VII. No obvious reason can be given for this variation. Sample pair No. IV were identical soybean oil samples; however 2 panels out of 12 report a significant difference between them. The results of this particular pair of samples show however that if the panel membership is large enough, the average scores have a small error. From a group of 134 tasters a difference in scores of only 0.1 unit was obtained. For a panel group of this size it would take a difference three times as large in order to indicate a significant difference between the samples. To reduce the error in this type of test it must be concluded that the consistency and discriminatory power of the individual taster must be considerably improved or that the number of members serving on a panel must be enlarged. Insofar as our results can be interpreted, panel enlargement or increasing the number of judgments rendered by the panel would be the most satisfactory solution. Dr. Schlosberg concluded from a similar study (56) ". . . expertness cannot take the place of a large number of judgments whether we consider the single expert or the expert panel."

TABLE VII

Agreement of Panels in the Detection of Differences Between Paired Samples

Sample pair		Total no. of tasters (all panels)	Score[a]	No. of panels reporting differences		
				No difference	Significant difference, 5% level	Highly significant 1% level
I	a	139	7.9	4	2	6
	b		6.3			
II	a	150	5.6	10	0	2
	b		6.2			
III	a	157	6.3	3	2	7
	b		4.6			
IV[b]	a	134	8.1	10	2	0
	b		8.0			
V	a	96	3.9	6	4	1
	b		5.2			

[a] Average of individual scores of all panels.
[b] Pair of identical samples.

REFERENCES

1. Anonymous, Taste Panels, Bibliography, Chicago, Quartermaster Food and Container Institute (mimeo.).
2. Beadle, B. W., Oil and Soap, *23*, 33–35 (1946).
3. Bengtsson, K., and Helm, E., Wallerstein Lab. Commun., *9*, 171–180 (1946).
4. Bliss, C. I., Anderson, E. O., and Marland, R. E., "A Technique for Testing Consumer Preferences, with Special Reference to the Constituents of Ice Cream," Storrs, Agricultural Experiment Station Bulletin 251 (1943).
5. Biggs, D. A., and Bryant, L. R., Can. J. Technol., *31*, 138–145 (1953).
6. Boggs, M. M., and Hanson, H., "Analysis of Foods by Sensory Difference Tests in Advances in Food Research," vol. 2, pp. 219–258, New York, Academic Press (1949).
7. Boggs, M. M., and Ward, A. C., Food Technol., *4*, 282–284 (1950).
8. Bradley, R. A., Biometrics, *9*, 22–38 (1953).
9. Bradley, R. A., and Duncan, D. B., Bi-Annual Report No. 1 on Statistical Methods for Sensory Difference Testing, Virginia Agricultural Experiment Station (1950).
10. Bradley, R. A., and Somerville, P. N., Progress Report No. 1 on Statistical Techniques for Sensory Testing of Processed Foods, Virginia Agricultural Experiment Station (1955).
11. Bradley, R. A., and Terry, M. E., Bi-Annual Report No. 2 on Statistical Methods for Sensory Difference Testing, Virginia Agricultural Experiment Station (1951).
12. Ibid. Bi-Annual Reports, Nos. 3 to 6 (1951–53).
13. Burrows, R. B., Food Manuf., *21*, 467–470 (1946).
14. Byer, A. J., and Abrams, D., Wallerstein Lab. Commun., *16*, 253–258 (1953).
15. Cairncross, S. E., Sjöström, L. B., Food Technol., *4*, 308–311 (1950).
16. Cartwright, L. G., Snell, C. T., and Kelley, P. H., Anal. Chem., *24*, 503–506 (1952).
17. Chang, S. S., and Kummerow, F. A., J. Am. Oil Chemists' Soc., *32*, 341–344 (1955).
18. Chirgwin, L. D. Jr., Eighth Annual Soybean Symposium, Chicago, National Soybean Processors Association, Soybean Research Council (mimeo.) (1952).
19. Dawson, E. H., and Harris, B. L., Agricultural Information Bulletin No. 34, USDA (1951).
20. Dunkley, W. L., Food Technol., *5*, 342–346 (1951).
21. Dutton, H. J., Schwab, A. W., Moser, Helen A., and Cowan, J. C., J. Am. Oil Chemists' Soc., *25*, 385–388 (1948).
22. Dutton, H. J., Lancaster, C. R., Evans, C. D., and Cowan, J. C., J. Am. Oil Chemists' Soc., *28*, 115–118 (1951).
23. Ehrenberg, A. C. S., Chemistry and Industry, *1950*, 93.
24. Evans, C. D., Lancaster, E. B., Dutton, H. J., and Moser, Helen A., J. Am. Oil Chemists' Soc., *28*, 118–121 (1951).
25. Grant, G. A., and Lips, H. J., Can. J. Res.. *F24*, 450–460 (1946).
26. Gridgeman, N. T., Food Technol., *9*, 148–150 (1955).
27. Handschumaker, E., J. Am. Oil Chemists' Soc., *25*, 54–56 (1948).
28. Hanson, H. Food Acceptance Testing Methodology, p. 57, Chicago Quartermaster Food and Container Institute (1954).
29. Harrison, S., Elder, L. W., Food Technol., *4*, 434–439 (1950).
30. Helm, E., and Trolle, B., Wallerstein Lab. Commun., *9*, 181–194 (1946).
31. Henick, A. S., Benca, M. F., and Mitchell, J. H. Jr., J. Am. Oil Chemists' Soc., *31*, 88–91 (1954).
32. Hopkins, J. W., Biometrics, *6*, 1–16 (1950).
33. Jakobsen, F., Food Technol., *3*, 252–254 (1949).
34. Jennings, W. G., Dunkley, W. L., and Reiber, H. G., Food Research, *20*, 13–22 (1955).
35. Kawahara, F. K., and Dutton, H. J., J. Am. Oil Chemists' Soc., *29*, 372–377 (1952).
36. Kenaston, C. B., Wilbur, K. M., Ottolenghi, A., and Bernheim, F., J. Am. Oil Chemists' Soc., *32*, 33–35 (1955).
37. Lappin, G. R., and Clark, L. C., Anal. Chem., *23*, 541–542 (1951).
38. Lips, H. J., and Grace, N. H., Can. J. Research, *F27*, 311–317 (1949).
39. Lockhart, E. E., Food Technol., *5*, 428–431 (1951).
40. MacLean, J. A. R., and Wickens, R.; Nature, *168*, 434 (1951).
41. Mackey, A. O., and Jones, P., Food Technol., *8*, 527–530 (1954).
42. Martin, C. H., Schepartz, A. I., and Daubert, B. F., J. Am. Oil Chemists' Soc., *25*, 113–117 (1948).
43. Moncrieff, R. W., "The Chemical Senses," p. 78, New York, John Wiley and Sons (1946).
44. Morse, R. L. D., "Rationale for Studies of Consumer Food Preferences, Advances in Food Research," vol. 13, pp. 385–427, New York, Academic Press (1951).
45. Moser, Helen A., Dutton, H. J., Evans, C. D., and Cowan, J. C., Food Technol., *4*, 105–109 (1950).
46. Moser, Helen A., Evans, C. D., and Cowan, J. C., Tenth Annual Soybean Oil Symposium, Chicago, National Soybean Processors Association, Soybean Research Council (mimeo.) (1955).
47. Moser, Helen A., Jaeger, Carol M., Cowan, J. C., and Dutton, H. J., J. Am. Oil Chemists' Soc., *24*, 291–296 (1947).
48. Neubert, A. M., and Carter, G. H., Food Technol., in press (1955).
49. Page, A. B. P., and Lubatti, O. F., Chemistry and Industry, *1948*, 723–727.
50. Pantaleoni, R., Chem. Eng. News, *31*, 1730–1734 (1953).
51. Peryam, D. R., and Swartz, V. W., Food Technol., *4*, 390–395 (1950).
52. Peterson, H. M., Eighth Annual Soybean Symposium, Chicago, National Soybean Processors Association, Soybean Research Council (mimeo.) (1952).
53. Pfaffmann, C., "Taste and Smell," in "Handbook of Experimental Psychology," ch. 29, Stanley S. Stevens, editor, New York, John Wiley and Sons (1951).
54. Pfaffmann, C., Schlosberg, H., and Cornsweet, J., Food Acceptance Testing Methodology, pp. 4–17, Chicago, Quartermaster Food and Container Institute (1954).
55. Schepartz, A. I., and Daubert, B. F., J. Am. Oil Chemists' Soc., *27*, 367–373 (1950).
56. Schlosberg, H., Pfaffmann, C., and Cornsweet, J., Food Acceptance Testing Methodology, pp. 45–54, Chicago, Quartermaster Food and Container Institute (1954).
57. Sidwell, C. G., Salwin, H., Milada, B., and Mitchell, J. H. Jr., J. Am. Oil Chemists' Soc., *31*, 603–606 (1954).
58. Sidwell, C. G., Salwin, H., and Mitchell, J. H. Jr., J. Am. Oil Chemists' Soc., *32*, 13–16 (1955).
59. Snedecor, G. W., "Statistical Methods," Ames, Iowa State College Press (1946).
60. Stapf, R. J., and Daubert, B. F., J. Am. Oil Chemists' Soc., *27*, 374–377 (1950).
61. Terry, M. E., Bradley, R. A., and Davis, L. L., Food Technol., *6*, 250–254 (1952).
62. Turner, E. W., Paynter, W. D., Montie, E. J., Bessert, M. W., Struck, G. M., and Olson, F. C., Food Technol., *8*, 326–330 (1954).
63. Wood E. C., J. Soc. Chem. Ind. (London), *68*, 128–131 (1949).