AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

Ag84Ag
1Cop.4

SENSORY METHODS FOR MEASURING DIFFERENCES IN FOOD QUALITY

Review of Literature
and Proceedings
of Conference

LIBRARY
CURRENT SERIAL RECORD
SEP 3 1951
U.S. DEPARTMENT OF AGRICULTURE

Agriculture Information Bulletin No. 34
Bureau of Human Nutrition and Home Economics
UNITED STATES DEPARTMENT OF AGRICULTURE

SENSORY METHODS FOR MEASURING DIFFERENCES IN FOOD QUALITY

Review of Literature and Proceedings of Conference

Prepared by

ELSIE H. DAWSON AND BETSY L. HARRIS

with the assistance of

Ruth A. Redstrom, Suzanne S. Alexander, Jessie C. Lamb, Ann F. Doyle, and Irene H. Wolgamot

BUREAU OF HUMAN NUTRITION AND HOME ECONOMICS

AGRICULTURAL RESEARCH ADMINISTRATION

Agriculture Information Bulletin No. 34

United States Department of Agriculture

Washington, D. C. August 1951

For sale by the Superintendent of Documents, U. S. Government Printing Office
Washington 25, D. C., Price 35 cents

Contents

Sensory Methods for Measuring Differences in Food Quality[1]

Food researchers must often rely on the senses of taste, smell, sight, or feel in food-quality evaluation. But so varied are laboratory methods of selecting food-judging panels, preparing food samples, setting up rating scales, and analyzing statistical findings, that the work of one group may not check that of another. The problem has become more acute with an increase in cooperative research in food quality.

Because of the need for standardizing procedures in taste-testing work to make possible comparison of results from various research institutions, the Bureau of Human Nutrition and Home Economics sponsored a conference for the critical appraisal of sensory methods for measuring food quality by taste panels. The 3-day meeting, held in Washington, D. C., January 23–25, 1950, was attended by representatives from many fields of laboratory research in which food-quality evaluation is a problem. Participants included home economists, food technologists, chemists, biologists, bacteriologists, horticulturists, plant physiologists, and statisticians. (See page 132 for list of participants.)

The scope of the discussions was limited to subjects relating to sensory methods used in the laboratory to measure differences in quality of food samples that have had different treatments. The conference did not cover problems of the market analyst who uses untrained taste-testing panels to forecast consumer acceptance of a product.

The proceedings of the conference are reported in this bulletin. The discussion under each subject is followed by a committee report which cites techniques thought to be desirable, and makes recommendations as to needed research on methodology.

In preparation for the conference, members of the Bureau staff made a careful review of the literature on methods of evaluating palatability. The review is included in this publication to give a more complete picture of the present status of palatability procedures.

REVIEW OF LITERATURE

This review of the literature on sensory methods for measuring differences in food quality, includes 300 selected references. The subject matter is arranged under the same major topics as those discussed at the conference. Appropriate subheadings have been added, and when the quantity of material under a particular heading is large, the information is further classified by commodities — beverages, cereal products, dairy products, egg products, fats and oils, fruits, meats, miscellaneous foods, poultry, vegetables — and primary tastes.

[1] Work carried on in part with funds alloted under the Research and Marketing Act of 1946.

Methods of Measuring Differences in Food Quality

Descriptive terms

Descriptive terms have been used in grading cereal products *(125)*[2], dairy products *(72, 132)*, eggs *(122)*, fats and oils *(13, 45, 146, 168)*, fruits *(54)*, meats *(9, 242)*, vegetables *(43, 57, 108, 115, 254, 266, 299)*, and primary tastes *(26, 236, 237, 238)*. Judges' comments are encouraged *(182)* and, in addition to giving numerical scores, judges have been asked to define the differences they noted *(103)*.

Dairy Products. Five quality groups were used (excellent, good, fair, poor, and bad) with 18 descriptive flavor defects. If the descriptive term given did not properly describe the defect, the judges used their own terms or said "unidentified." This was done for odor also *(288)*.

Eggs. In addition to using numerical terms, the judges were instructed to give their comments on the type of flavor and any departure from normal flavor *(19, 259)*.

Meats. Reasons were given, where possible, for preferences after samples were ranked *(61, 294)*. In the first year's study of beef, only descriptive terms were used *(24)*.

Miscellaneous foods. Frozen precooked meals: When a score was low, the judge was requested to give reasons for it *(126)*.

Poultry. The use of descriptive terms, even though worked out by the panel of judges, was not of significance *(286)*. The fact that judges were asked to report off-flavors undoubtedly influenced some to ascribe off-flavors to samples they would have considered good under ordinary circumstances *(12)*. However, descriptive terms have been used to describe flavor *(150)*. In a turkey experiment judges checked whether there was a strong, medium, slight, or no fishy odor and flavor *(192)*.

Vegetables. In addition to scoring, adjectives to describe each factor were listed in columns and judges were instructed to check those that best applied to each sample *(231)*. Descriptive terms such as excellent, good, fair, poor, or unpalatable were used *(266)*. After preference rating had been made on samples of corn, judges gave a written statement for reasons of placement *(101)*. After scoring in another test, each judge was asked to state how he liked his vegetables, that is, the degree of doneness he preferred *(114)*. Judges' comments were requested in addition to paired judging that was done *(290)*. After scoring, judges noted whether they detected an off-flavor or taste of sulfite *(277)*.

Primary tastes. In addition to numerical scoring, taste was identified *(166)*. Some terms used were: Tasteless, sweet, bitter, sour, salt, or combinations of the four basic tastes *(27)*.

Numerical scores

Numerical scoring has been employed to evaluate either general quality or a maximum of two specific characteristics *(182)*, and to record differences along with positive and negative attitudes of the judges toward the differences *(65, 102, 227a)*.

[2] Italic numbers in parentheses refer to Literature Cited, p. 18.

The fundamental supposition of any rational quality grading is that the number expressing the grade is proportional to the quality of the property to be measured *(228)*. The weight given to the grade should be determined by preliminary testing *(64)*. An equal difference in grade numbers must correspond with an equal difference in quality. As two successive grades always differ by one unit, the terms used to describe two successive qualities should be selected so as to express equal differences in sensation. The relationship between terms and grade numbers is illustrated in the chart below. Its grading system allows a ±5-percent range of tolerance *(228)*.

Quality term		Quality range in percent	Grade
First alternative	Second alternative		
Perfect (fancy)	Excellent	Over 95 percent	10
Excellent	Very good	85 to 95 percent	9
Very good	Good	75 to 85 percent	8
Good	Slightly good	65 to 75 percent	7
Slightly good	Borderline plus	55 to 65 percent	6
Average (mean)	Borderline minus	45 to 55 percent	5
Fair	Slightly poor	35 to 45 percent	4
Borderline	Poor	25 to 35 percent	3
Bad (defective)	Very poor	15 to 25 percent	2
Very bad	Extremely poor	5 to 15 percent	1
Unacceptable (not eatable)		Below 5 percent	0

It is unsuitable to use different grade numbers for the same descriptive term in different factors, but such differences in numbers for the same descriptive term become unavoidable if different "best points" are introduced for different factors *(228)*.

A scale is an arrangement of perceptibly different quantities of a property in a graduated series. It may be imaginary, that is, used with descriptive terms, or it may be material, having specimens of the product or other substances chosen for direct comparison with the experimental product *(269)*.

The number of gradations in the grading chart is important in accuracy and application *(269)*. It will depend on the number of intervals that the judge can distinguish and it may be necessary to make allowance for the fact that judges may not use the highest and lowest scores *(37)*. Uniformity of opinion is likely to increase with a decrease in number of such grades, but if the scale is too coarse, it will not have much scientific value *(269)*. Probably 10 intervals would be sufficient in most experiments *(37)*. If the operative range of scores were narrow in a product, even poor judges might make a good showing in that product. However, a relatively good judge might appear, on the basis of closeness of scores, to be doing poor judging if the operative range of scores were wide *(282)*.

Using 10 grades, it is sufficient to grade partial food properties in integer numbers and not to introduce decimal fractions. The advantages of this method of grading are that it is uniform and simple, and that quality is expressed in descriptive terms that can be selected so as not to contradict appreciably the desired proportionality. Whether the scale should be restricted to 5 or 10 grades, or extended to 15, 25, or even 100 depends on the sense-thresholds of the tasters. Ten integer numbers

should be sufficient for the quality determination of any single property. The total grade expressed in terms of such single property grades may then contain decimal fractions *(228)*.

A 10-point scale has been used by other investigators *(103, 152, 162, 193)* where there were 5 equal degrees of acceptability and 5 equal degrees of nonacceptability *(103)*; where a 1–10 range was used with 10 equal to excellent *(152)*; or where 0 equaled unacceptable and 10 equaled excellent *(65, 162)*. A scale in which 10 equals excellent, 9, 8, 7 equal good, 6, 5, equal fair, 4, 3 equal poor, and 2, 1 equal very poor gives uniformity in scoring and is helpful in analyzing the results *(153)*. On the other hand, it is considered by Bate-Smith *(18)* that a scale from 0 to 10 may easily be heterogeneous. The range 8–10 may involve no unpleasantness, whereas the range 0–8 may *(148)*. The 10-point scale lends itself easily to subjective evaluation of quality of widely varying types of foods *(162)*.

The advantage of the use of a 5-point scale, with only the highest and the lowest points defined is that it avoids much of the difficulty of devising adequate description of flavors; it is difficult to achieve linearity when every point is defined *(154)*. In using a 7-point scale with descriptive terms for each point, it is not necessary to depend on adequate definition of only the 2 anchor points, and the tasters can be directed how to assess the relative importance of different defects, for example, lack of flavor or presence of off-flavors *(154)*.

One of the main difficulties with regard to the method of assessing total food quality by point scoring is that one characteristic may render the food totally inedible and yet the total score may be relatively high. Although this discrepancy is reduced somewhat by weighting the scores, obviously it can never be entirely eliminated. The difficulty may be overcome, however, by taking the product of the individual scores (geometric summation) in place of the sum (arithmetic summation) *(131)*.

Cereal products. Numerical scores have been used in judging cereal products *(113, 130, 190, 204, 249)*. Pastry was judged with descriptive terms, and numerical values were assigned each rating: 5, excellent; 4, good; 3, fair; 2, poor; and 1, not edible *(217)*. Bread was judged on a 3-point scale with 1 indicating best, 2 representing intermediate, and 3 representing poorest *(159)*. Bread has also been judged on a 4-point scale: 4, excellent; 3, good; 2, fair; and 1, poor *(167)*. Ration biscuits were judged by the following 10–0 scale: 10, excellent; 8, good; 6, fair; 4, poor; 2, very poor; and 0, inedible *(196)*. Corn meal and macaroni products have been judged by a 1–7 scale with seven as most desirable *(218)*.

Cake was given a weighted scale totaling 100 points: 30 points allotted for crumb (texture), 20 for tenderness, 20 for velvetiness, and 30 for eating quality (aroma, flavor, and over-all) *(47, 197)*. In another cake experiment the official A. A. C. C. (American Association of Cereal Chemists) method of weighting was used: Symmetry, 10; volume, 15; crust, 5; texture, 30; grain, 25; color, 15 (*38*). Toast has been judged on a weighted scale allotting 20 points for color, 10 for character of crust, 15 for character of toasted surface, 15 for character of crumb, and 20 each for aroma and taste (*67*).

Dairy products. A standard and uniform terminology must be adopted before a well-founded technique of flavor scoring can be formulated *(164)*.

A 0–10 scale, with 10 as excellent, has been used to judge dry milk *(222, 224, 225)*. A 0–15 scale was considered by one investigator of dry milk to be too wide *(164)*.

Milk has been judged on a scale of 0–25 *(288)* developed by the U. S. Bureau of Dairy Industry; on a scale of 1–25 *(247)*; and according to the flavor-scoring system of the American Dairy Science Association *(92)*. A 0–3 scale, with 3 equal to excellent, has been used for judging milk *(224)*. A 0–4 scale has also been used, where 0 indicated no oxidized flavor; 1, slightly oxidized flavor; 2, oxidized flavor; 3, pronounced oxidized flavor; and 4, very pronounced oxidized flavor. Those samples marked 0 and 1 were considered salable and those receiving 3 and 4, not salable *(10)*. The oxidized flavor of milk, butter, and ice cream have been judged with the use of plus and minus signs *(74, 119, 283, 284)*. In one case, numbers from 1 to 6 indicated increased intensity of oxidized flavor *(74)* and in another, numbers 1 to 10 were assigned to the plus and minus values *(119)*. Weighted scores were given for ice cream *(105, 198)* and butter and cheese *(105)*.

A dry milk and egg mix was judged according to the following system: The standard was given an arbitrary score of 10; 20 indicated a product twice as acceptable, 5, a product half as acceptable, and 0, unacceptable *(226)*.

Butter was given a 0–10 scoring, where 10 was excellent; 8, good; 6, fair; 4, poor; 2, bad; and 0, inedible *(129)*. Cheese was judged on a 1–10 scale with 1 representing the softest cheese and 10 the firmest *(246)*, and on a scale whose range was 45 points *(281)*. Flavor standards for Canadian Cheddar cheese were given: First Grade had a minimum score of 39, Second Grade had a minimum score of 37, and Third Grade a score less than 37 *(155)*.

Eggs. A scale of 0–10, with 10 as excellent, has been used in judging eggs *(156, 273, 274)*. Scales of 1–10, with 10 as excellent *(3, 98, 99, 141, 170, 259, 296)*, and scales of 0–8 *(18, 73, 180)* have also been used. One to five scales have been used, in some cases with 5 representing best quality *(98, 253, 271)*, and in others with 1 as best *(187, 248)*. Letters were used to judge eggs, and were converted later to numerical grades: F, indicating fresh or perfection, 4 points; A, indicating excellent, 3 points; B, indicating good, 2 points; C, indicating edible, 1 point *(227)*. Weighted scores totaling 100 have also been used *(205, 267)*.

The rating for consistency of eggs by a 5-point scale was less reliable than the rating for flavor of the same eggs by a 10-point scale, perhaps because of the smaller number of choices in the rating scale *(98)*. The 0–8 range was considered too wide to lend itself to the best in statistical analysis. Most of the egg samples scored from 6 to 8; the widest variation between tasters occurred on low-scoring samples *(180)*.

Fats and oils. Numerical scores and descriptive terms have been used to judge fats and oils *(80, 104, 213, 245)*. A scale of 0–10 was used in judging salad oils and lard, with 10 as top quality or excellent and 0 indicating poorest quality or unapproachable *(89, 128)*.

Fruits. Rating scales with a 1–7 range and a 1–5 range were used in scoring different fruits *(17, 136, 137, 165)*. Strawberries were judged by a weighted scale totaling 18 points *(127)*. Citrus fruits have been rated on a 20–100 scale with 20-point intervals *(143, 144)*.

Meat and poultry. Wide use has been made of the 1–7 grading

chart which was adopted by the Cooperative Meat Investigations for judging meats and poultry *(14, 23, 24, 40, 42, 68, 70, 107, 138, 147, 150, 186, 216, 252)*. Other kinds of numerical scores used in scoring meats include: A 1–5 range where 5 represents the greatest desirability *(220)*; 0–10 where the larger values indicate greater preference *(221, 222, 223, 260)*; 0–6 range for no change to extreme change in flavor of canned meats (*123*); and tenderness ratings given values of 2, 4, 6, up to 14 (*234*).

In judging poultry, a 0–10 range of scoring, with 10 as perfect *(117, 185, 263, 264)* has been used, and 1–10 with 10 the highest score *(142, 295)*. When any portion of the fowl was described as inedible because of its moldy condition and rank odor, and not cooked for scoring, it was automatically given a score of zero for aroma and flavor *(295)*. A 1–5 range has been used for poultry, with 5 being perfect or excellent *(171, 278)*. Scoring on a 1–4 scale has also been done, with 1 indicating good and 4 indicating strongly off flavor *(12)*. Odor of poultry has been judged with a 0–3 scoring, where 0 indicated no foreign odor, and 3 an intense and disagreeable odor *(240)*. Aroma and flavor have also been scored on a range of 50 *(244)*.

In judging meat, weighted adjectives arbitrarily set at 1 to 5, with 5 as very tender, have been used in addition to paired eating tests and found to furnish a valuable supplement *(75, 76, 77, 79)*. Numerical values can be given to preference ratings *(61, 294)*. Other types of weightings give separate characteristics a number of marks which make up a maximum or total of 100 *(157, 174)* or a total of 49 (7 for each of 7 factors) *(127)*.

Vegetables. Seven-point scales with 7 as optimum and 5-point scales with 5 as optimum have been used in scoring vegetables *(15, 114, 135, 231, 261, 285)*; also 4-point scales with 4 as excellent *(270, 277, 290)*, 4-point scale with 1 as highest score *(46)*, and a 0–4 scale with marks of ½, ¼, and ⅛ *(276)*. Possibly a scale of 0–10 with no fractional marking allowed would have been more convenient, for tasters frequently find reasons for departing from any suggested marking system *(276)*. One 3-point scale made use of the numbers 1, 3, 5, to denote poor, fair, and good, respectively *(172)*.

Color, flavor, texture, and general acceptability are often rated on a 10-point scale *(60, 97, 255)*, but the interpretation on the scale may vary greatly:

(1) Where 1 and 2 equal highest excellence; 5, fair; and higher than 5, progressively poorer to undesirable *(50, 52)*.

(2) Where 1–2 is excellent; 2–3.5, very good; 3.5–5, fair to good; and 5–10, progressively poorer *(49)*.

(3) Where 1–2 is excellent; 2.1–3.5, very good; 3.6–4.5, fair; 4.6–5, poor; and higher, very poor. In addition, each judge summed up each sample in a rating on general desirability (37½ points each for flavor and color and 25 points for texture) (*91*).

(4) Where 1–1.5 is best; 1.6–2.5, very good; 2.6–3.5, good; 3.6–4.5, fair; and higher than 4.5 is poor to very poor *(53)*.

(5) Where 1–1.9 is excellent; 2–2.5, very good; 2.6–3.5, good; 3.6–5, fair; 5–6.5, poor; and 6.5 and above is very poor *(298)*.

(6) Where 1–2 is highest score, 2.1–3.5 is very good, 3.6–5 is fair to good and higher than 5 is poorer. In addition to this, each sample is given a score upon comparative desirability as a food product *(51)*.

An A-B-C score has been used to evaluate marketability, where A equals highest market quality and C too tough to be marketed *(50)*. Another type of numerical rating was as follows: 9, 8, 7 equal high market quality; 6, 5, 4 equal marketable but tougher; and 3, 2, 1 equal not marketable *(34)*.

A weighted scale totaling 24 possible points was used for judging snap beans, and weights totaling a maximum of 27 points for peas *(127)*. In another experiment on peas, weights added up to a maximum of 100 points; judges also counted the number of hard peas in the sample and one-third of the percent of hard peas was subtracted from the total score *(30)*.

In one vegetable experiment, the equal sign (=) meant equal to the standard; +2, +4, +6 indicated degrees superior to the standard, and −2, −4, and −6, degrees inferior to the standard *(44)*.

The highest score for cucumbers was 9 for texture, 18 for flavor, and 18 for palatability *(292)*. One investigator used weighted scores which totaled a maximum of 100 points *(287)*. Weighted scores, 20 points for color, 40 each for texture and flavor were used; judges were instructed to rate sample they thought best in each factor 100, and grade the others from that standard *(55)*.

After samples of potatoes were ranked, tasters indicated the presence or absence of off-flavor by plus and minus signs, the number of minus signs indicating the intensity of off-flavor *(134)*.

Frozen precooked meals have been judged on a scale of 0–100 with intervals of 20, 100 equaling a perfect score *(126)*. Primary tastes have been scored numerically on a 0–5 scale, where 0 indicated no taste and 5 indicated very strong taste *(166)*.

Ranking tests

In the ranking test judges are asked to rank samples in decreasing or increasing order of some characteristic *(37)*. Ranking tests have been made on all types of food products, the numbers of samples to be ranked ranging from 2 to 10 *(22, 31, 50, 58, 61, 65, 72, 82, 89, 95, 101, 133, 134, 163, 167, 202, 210, 217, 251, 258, 271, 284, 294)*.

An arrangement for the exercise of vision, scent, and taste, separately and jointly, on the same series of samples was effected in ranking. A standard (the sample of lowest quality) was designated as unity, and the others were ranked upward from it as a fixed base. First, the judge ranked the series by sight alone, then (blindfolded) by taste alone, and finally by combined taste and sight. Another experiment was designed to test the sensitivity of the sense of smell. The judge was blindfolded and made duplicated tests by (a) scent alone, (b) taste alone, (c) scent and taste together, and (d) unblindfolded, by sight, scent, and taste in combination *(82)*.

Paired tests

In the paired test two samples are submitted to judges *(37, 77)*. The paired samples are judged by comparison with each other *(77)*. A typical question asked of the judges is "Which is the more tender sample of meat?" Sometimes a standard sample is presented first, and the judges are asked which of the two unknowns is the same as the standard *(37)*.

Every person asked to judge will express a choice even though he has no real choice between the two samples, and a retest at a later date may

reverse the first judgment. It is best to retest the same group to determine the number of these guesses, and since the law of probability will indicate that half of the guesses will be reversed, the number of such guesses should be multiplied by two. Final results will show: (1) Favorable reaction, (2) unfavorable reaction, (3) the number of those who have no real preference *(1)*.

This test limits the number of samples to two and, for any large number of treatments, a comparison of only two at a time is very costly *(37)*, and unwieldy *(202)*. On the other hand, the paired method permitted the judging of a larger number of eggs without tiring the judges, because few standards had to be remembered when only two eggs were being compared at one time *(122)*.

Use of paired tests has been reported as a method of judging the palatability of various foods *(1, 103, 152, 153, 154, 173, 202, 227a)*, cereal products *(66)*, dairy products *(62, 164)*, eggs *(96, 122)*, fats and oils *(104, 146, 213)*, apples *(17)*, vegetables *(290)* and precooked meals *(126)*.

The paired test has been used extensively in meat experiments *(75, 76, 77, 78, 79, 138, 174, 186, 195, 233)*. One investigator reported the use of paired bites from paired slices from paired roasts *(76)*. The paired-eating method for testing tenderness has the advantage of direct comparison of two paired samples and proved to be satisfactory for testing differences in tenderness of meat resulting from two methods of cooking. It is suggested also as a method suitable to use in perfecting objective methods. It is applicable only in those cases in which the difference between two comparable samples is considered. It cannot be used for comparing a large number of individual roasts or roasts cooked on different days *(75, 77)*.

Triangle or triple comparison tests

In the triangle test, three samples are examined, two of which are duplicates. Judges are asked whether there is any difference between samples and, if so, to select the identical samples. Often judges are also asked to indicate whether the odd or duplicate samples have the distinguishing characteristics to the more pronounced degree *(37, 153, 239)*, or which sample is preferred *(151)*.

The test can be used first with an expert panel to determine whether a difference exists and then with a large number of people to determine consumer acceptability of either or both samples *(239)*. It may also be employed in selecting personnel for expert panels, wherein individual sensitiveness to different taste factors are evaluated *(22, 152, 153, 239)*. It also lends itself to statistical analysis *(239)*.

If one is asked to select the two similar samples from a set of three, it is possible to make three different selections, only one of which is correct. Chance selection alone will give one correct answer in every three trials (33 percent). The percentage might be altered somewhat in one direction or the other by circumstances which would not be equalized in a small number of tests. To determine significance of the results, it is necessary to know how far the number of correct answers must exceed 33 percent before it may be considered certain that guessing has been eliminated *(151)*.

There are several references to the use of triangle tests *(44, 58, 152, 153, 188, 227a, 239)*. This method is found to be very satisfactory and highly recommended *(153)*.

Dilution tests

The dilution test determines the smallest amount of unknown that can be detected when it is mixed with a standard material. It applies only to homogeneous substances, but many foods can be made homogeneous without effect on flavor. The dilution method as it has been applied probably has given the most accurate results of any of the methods. This test could be used in any of the methods for flavor and odor testing. It is especially well adapted to the difficult problem of storage studies for which all other approaches have shortcomings. However, the lack of suitable standard material will prevent its use with many foods *(37)*.

Perhaps the most popular usage of the dilution technique is in the primary taste tests *(29, 84, 95, 112, 166, 169, 236, 241, 243, 247, 251, 297)*. Beakers containing dilutions of a primary solution were shuffled and the judges were instructed to place the solutions, to the best of their ability, in the order of their concentrations. From the standpoint of individual selection alone, a series of dilutions should contain as large a number of samples as possible within the limit which may be handled expeditiously *(284)*. Other series were arranged in order of sweetness *(59)*.

A series of 10 samples of oxidized milk was used in studying the effect of temperature on accuracy of judgment. The amount of oxidized milk, which was added to fresh milk, was decreased approximately 10 percent for each sample *(284)*. The percentage dilution of six different dilutions of dried egg in fresh scrambled egg was based on a logarithmic scale *(39)*. Potato samples were placed in order of increasing dosage of insecticide. Tasters were asked to check the presence or absence of off-flavor *(134)*.

Difference preferance tests

Difference preference tests are useful in detecting differences, and in determining which difference is preferred *(103)*. Preference rating is considered the simplest method of judging, the most applicable for consumer surveys, and a valuable aid in selecting a panel. Also it is a good method to use when only a few samples are given and when the difference between the samples is only slight *(65)*. However, preference data are not always considered permanently of predictive value; the only dependable method of ascertaining population preference is a quite complete survey *(202)*.

Constant stimulus differences method

The constant stimulus differences method is used, in which two stimuli are presented and the individual is told to state whether the second stimulus is more or less intense than the first. Random order is necessary for this method since the second will be judged greater than the first when the two are of equal intensity *(202)*.

Matching with standards

In testing coffee the sample of unknown freshness was compared with each of a standard series made up of varying quantities of fresh and stale coffee *(232)*. For fats and oils five control samples and one test sample were compared. The judges were requested to rate the samples in order of increasing concentration of reversion odor. When a judge misplaced the controls in ranking, his scores were discarded *(140)*.

In one case, solutions to be tested were tasted against a selected standard until a concentration was found which possessed sweetness that compared in intensity with that of the standard solution *(93)*.

Judges matched molar solutions where the solution to be matched contained a greater than taste-threshold concentration of a substance and, in addition, a subtaste-threshold concentration but greater than sensitivity-threshold concentration of a contrasting substance. This method was very satisfactory in determining the effect of subtaste-threshold concentrations of one substance upon mildly strong-tasting concentrations of a contrasting substance *(112)*.

Other methods of testing

(1) Solutions were tested at a large dinner, where the people present rated the solutions by raising their hands. It was impossible to obtain an accurate record with a test of this kind *(26)*.

(2) Reaction of panel was judged from the amount left on the plates. General appearance, color, flavor, texture uniformity, and defects were also noted *(275)*.

(3) Judges were instructed to rate the sample 100 if they thought it best in color (20 points), texture (40 points), and flavor (40 points), and grade the other samples from that standard *(55, 56)*. This method of grading for each factor against a visible standard gave close agreement between the scores of different judges *(55)*.

Comparison of different tests

The use of ranks rather than numerical scores encourages the judges to make fewer distinctions among samples. It has the advantage of reducing the tendency of individual judges to prefer certain score ranges *(217)*. Ranking system has the distinct advantage of simplicity over a numerical grading system. An individual must have considerable training before he can obtain consistent results in grading *(31)*.

Numerical scores with descriptive terms are better than numerical scores alone because the latter method does not distinguish the degree of difference between best and poorest *(296)*.

Judges indicated that they preferred a score sheet with numerical scores and descriptive terms to the vertical line score sheet, but more significant statistical results were obtained from the latter *(17)*.

Type of score card used

Most score cards list the appropriate characteristics to be judged, with descriptive terms or numerical ratings pertaining to each. Careful planning is necessary to determine the proper method of obtaining a score for the food. Scores obtained from preliminary panel testing should be analyzed, and if necessary, the factors discussed and changed before further scoring takes place *(63)*. Uniformity and simplicity are the characteristics of a rational grading system *(228)*. For trained organoleptic panels, printed directions should be brief *(65)*.

To score apples, a grading chart was used which consisted of a horizontal line 6 inches long. The left end of the line, representing "very poor," was considered to have a value of 0 and the right-hand end, representing "excellent," a value of 6. Judges were asked to place a short vertical line across the horizontal line where, in their opinion, the sample should be

classified. These points were then given values equal to the distance from the zero point *(17)*.

Odor detection

Many means for the detection of odor have been suggested. Crocker and Sjöström *(88)* have grouped them into physical means, chemical means, and odor accumulators. Included in the physical means for detecting odor are (1) electromagnetic radiation, (2) spectrograph, (3) electronic Halogen detector of the General Electric Company, and (4) hygrometry. Chemical means include (1) the stinkometer, (2) the use of carbon monoxide passed through silver permanganate, and (3) the Gutzeit test. Odor accumulator substances such as water, glycerol, a bland oil or absorptive carbon can be used. Moncrieff *(211)* suggests using the stinkometer to measure the reducing volatile matter in foodstuffs to detect incipient spoilage before any change is perceptible by the sense of smell. Crocker and Sjöström *(88)* say that the stinkometer has no applicability to odors in general, especially those of only "smelling strength."

One of the personal measurements that has been devised is the blast injection test *(85, 109, 110)*. At regular intervals, an amount of air and odor is injected into the nasal passages until the number of cubic centimeters necessary to identify the odor is ascertained. Such substances as coffee, citral, oil of turpentine, and benzaldehyde are most advantageous for this type of test *(109)*.

Showalter *(251)* recommends the use of filter paper in smelling tests. A few drops of the sample are put on a strip of filter paper and smelled by the judges. For smelling bottled products, he suggests using an "osmoscope," a glass tube which fits over the nose and extends down into the bottle.

The following chemical odor test for benzene hexachloride has been developed by the Beech-Nut Packing Company: The food is mixed well with benzene, the benzene is then poured off and filtered, and an aliquot is evaporated to dryness with a gentle current of air. After chilling in an ice bath, cold nitrating mixture is added, and the flask is placed in a boiling water bath for a half hour. The flask is then cooled and sodium carbonate solution plus sodium hydroxide is added. At this point, the odor of benzene hexachloride is detected if any is present.

Ford *(118)* conducted an investigation on odorous substances by allowing the judge to sniff the odor of an unknown substance which was in a covered bottle held by an assistant. Vail and Conrad *(286)* conducted an odor test on poultry using the following procedure: Individual birds cooked in covered containers were removed from the heat; each judge then smelled the bird, removing the cover only long enough to whiff the steam. About 15 minutes after removal from the oven, the birds were again tested for odor. Then covers were removed from the containers, the birds cooled for about 5 minutes and once more evaluated for odor.

The Crocker-Henderson system of odor analysis was devised to measure odor intensities. This system operates on three assumptions: (1) That the human nose is provided with four and only four kinds of odor nerves; (2) that every odoriferous substance, sniffed in adequate amount, stimulates all four kinds of these nerves simultaneously to the extent characteristic of the substance and of its concentration; and (3) that the odor sensation chord thus created by the excitation of nerve endings is capable of producing a distinctive odor impression *(86)*. By this system an odor may be represented as a four-digit number. The first digit is the measure

of relative fragrance, the second of acidity, the third of burntness, and the fourth of caprylic character. Each component in the odor in question is determined by comparison against the same component in the odors of the chemicals of a set of accepted standards *(7, 85)*.

Panel Selection

Experience

The art of tasting without prejudice can be acquired only by experience *(151)*. Panel members with previous experience and training are preferred *(22, 65, 161)*, and experienced tasters have been shown to obtain better results than the inexperienced *(151)*. The trained panel member has a knowledge of judging techniques and critical analysis *(227a)*, detects differences unheeded by the untrained *(22, 161, 207)*, describes better his taste impressions *(22, 207)*, is more accurate and reliable *(65)*, and has a better understanding of the terminology used *(22)*.

The human organism is our only proper recording device for flavor in its entirety; education of nose and tongue is the surest way of making it reliable *(83)*. There is considerable transfer of skill from one organoleptic problem to another *(65)*. However, there is no evidence that a judge who is sensitive to one flavor is equally sensitive to another flavor; therefore, panel members should be chosen on the basis of the characteristic to be evaluated in a particular study *(65, 102, 154, 227a)*, not on their performance on other panels *(227a)*. The panel need not be large, but its members should be experienced in the tests being made *(154)*. It is generally unsound for expert tasters to work subjectively on the basis of their own over-all preferences because their very expertness makes them atypical of the general consuming public (*89, 235*).

Most of the references reviewed reported the use of experienced judges with dairy products. One panel made up of both experienced and inexperienced members resulted in accurate and dependable judgments from the latter and superior performance from the experienced judges *(288)*.

If egg judges are inexperienced, their number should be greatly increased in order to get a reliable cross section of the public at large *(209)*. The importance of using the same panel of judges is stressed in connection with judging a series of egg samples for which maximum precision of relative assessment is desired *(156)*. But with oil samples, past experience appeared to have little value, as some individuals who had been grading oil for many years were frequently unable to arrange controls in correct order *(140)*.

Meat judges are needed with experience that covers a complete range of quality *(9)*; poultry ratings should be made by persons who can remember degrees of quality over long periods of time and consistently rate these degrees of palatability *(184)*. With tests on vegetables, emphasis was also placed on the importance of using the same panel each year *(285)*.

Work in judging primary tastes has been carried on with both experienced and inexperienced tasters. One worker reports that as the work proceeded, the tasters improved in delicacy of perception *(58)*. Series placement helped to develop skill in tasting through teaching tasters to look for and interpret signs rather than through increasing the sensitiveness to stimuli. The tasting of pure solutions of various concentrations

reveals some of the "signs" and thereby teaches the fundamental principles underlying taste judgments *(284)*.

Availability

Availability as a factor in selection of panel members is reported in connection with the judging of a variety of foods *(24, 43, 61, 114, 133, 134, 135, 150, 185, 186, 200, 250, 266, **272**, 276, 286, 291, 294, 299)*. Judges must be reasonably accessible *(207)*, and for this reason, they are preferably from the immediate staff *(188, **241**)* or from related staffs *(188)*.

Age

Age differences in receptors per unit area (for threshold tests) are a factor to be considered *(202)*. Preference as to desirable age ranges for judges vary greatly. A variety of ages is considered desirable by some authors *(207, 210)* as in egg judging *(156, 209)*, while one investigator of dairy products concluded that age was not found to be a factor in scoring *(288)*. Definite age flavor preferences were shown with certain foods *(6)*.

Discrimination of taste was believed to decrease with age *(69, 201)*, and olfactory powers appear dulled with age *(26)*. The preference for sweets declined in the oldest group while the preference for tart fruit tastes rose *(173)*.

One author thought that persons under 30 years of age appeared to have significantly lower taste sensitivity *(156)*; another considered the optimum age range as 30 to 40 years *(22)*; and one believed maximum sensitivity to lie in the age group 30 to 39 *(151)*. Preference curves for the group 12 to 18 years of age closely paralleled the curves for the age group 20 to 40 years *(173)*.

In testing primary tastes both homogeneous and mixed age groups have been used, including children beginning at 7 years and progressing to adults of 85 years *(166, 169, 175, 176, 236, 237, 241)*. One author concludes that there is probably no correlation between ability to identify primary tastes and age *(169)*, and another feels that age is not necessarily associated with poor tasting ability *(25)*.

Sex

Both sexes are used in many judging panels *(5, 188, 207, 227a, 229)*. Taste deficiency is primarily due to a single recessive gene, not sex-linked nor sex-influenced. When neither parent can taste the compound, none of the children can *(256)*.

One investigator reports that females had significantly lower taste sensitivity *(156)*; another that men tend to excel in identifying solutions by taste while women excel in identifying odor *(289)*.

In judging dairy products, differences in flavor preferences between the sexes were not large enough to be considered *(31)*.

Conclusions from a consumer preference test on primary tastes *(175)* were:

(1) Over 50 percent of both men and women preferred moderately sweet and salty foods.

(2) More women than men preferred excessively salty and sour foods.

(3) Over 50 percent of the men liked slightly sour foods.

(4) Women showed greater sensitivity of taste in distinguishing between the four basic tastes: Sweet, salt, sour, and bitter.

Health

Panel judges should be persons of good health and appetite, not susceptible to mouth and sinus infections *(227a)* or to colds *(69, 201, 227a)*. Colds affect the sense of smell *(28)*. Judges should be physically well, not fatigued or worried *(188)*.

For the judging of primary tastes, persons were eliminated who had numerous head colds *(95, 166)*, mouth or sinus infections, or allergies to a large number of foodstuffs *(95)*. No one was tested while he had a cold *(95)*.

Psychological factors

Successful conduct of taste panels is frequently as much a matter of human relations as a scientific problem. Panel members must have a keen interest in their tasting ability and these feelings must be sustained. Informal conferences should be held periodically *(213)* and imagination and suggestion must be eliminated *(201)*. Other psychological factors should be considered, for example, beer should not be tasted from a cup or tea from a glass *(82)*.

Panel members judging oil should have an interest in oil problems in general, and a desire to participate *(213)*.

There is no close correlation in testing of primary tastes between thresholds and emotional response *(25)*. Sensitivity of taste is found under conditions of repose and freedom from distractions *(208)*.

Concentration of odor has much to do with our likes and dislikes, and odors are also tied up with associations that make them pleasant or unpleasant *(26)*.

Assessments of the palatability of foodstuffs depend upon olfactory and tactile as well as upon gustatory sensations and are further conditioned by the subjective reaction of individuals to these stimuli *(156)*.

Taste and smell sensitivity

The selection of judges with high sensitivity of taste and olfactory sense is suggested by some authors *(69, 154, 201, 206)*. Important, too, is the ability to judge one feature at a time and, within limits, to disregard saltiness, sweetness, acidity, bitterness, or other distractions *(295)*. Special tests should be arranged to detect those individuals who are most capable of recognizing differences in taste. As a rule, the "triangular" test should be employed. When the most sensitive tasters have been selected by this method, their efficiency should be checked regularly by a study of the data from the routine tests *(22)*.

Nine of the references reviewed reported the use of taste-sensitivity tests *(65, 69, 124, 188, 201, 202, 206, 207, 227a)*. Some made the tests with pure solutions *(69, 201, 202, 206, 227a)* and others used the food in question *(65, 188, 227a)*. Three of the references reviewed reported the use of smell-sensitivity tests *(65, 206, 227a)*.

Variations in thresholds may be caused by unequal familiarity with substances used, also by degree of hunger, preference value, diet, smoking, and several physiological conditions. The area of sense receptors stimulated must be constant *(28)*. Individual reactions to taste depend primarily on innate hereditary factors and environment is of little importance. Smell also depends on hereditary factors, but environment is of more importance *(28)*.

In order to make sure that judges are suitable for testing work, they should be tested with the food in question (cereal products) for taste and smell sensitivity, strength, and desirability *(191)*. In judging baked products, judges selected on the basis of taste-sensitivity tests with pure solutions did not differ greatly in their ratings from the others *(66)*. Capable judges include only those whose senses of smell and taste are keen *(159)*.

In judging dairy products, it is important to calibrate judges by threshold tests before a dependable scoring method can be estimated *(164)*. Egg judges should be sensitive to the known off-flavor and respond quantitatively to variations in their off-flavor *(4)*.

Panel members for fats and oils were given taste-sensitivity tests *(80, 104, 146, 213)* and smell-sensitivity tests *(80, 104, 140, 146, 213)*. Smell-sensitivity tests with reverted soybean oil were given as follows: Five samples ranging from all-soybean oil to all-cottonseed oil in 25 percent steps were prepared and reverted. Prospective panel members, who had just been made acquainted with the reverted soybean odor, were asked to rate the samples in order of increasing reversion odor concentration (*140*).

There is no evidence that a direct relationship exists between sensitivity to the taste of chemically pure solutions and ability to detect flavors in food products *(95, 227a)*. Therefore, persons with high as well as average or low thresholds should be included in food judging panels in order to obtain more information.

Determination of sweetness, sourness, salinity, and bitterness is purely comparative. The quantity of solution tested, the temperature at which it is tasted, and even the time of day — all are factors influencing taste *(111)*.

The threshold concentration of primary taste substances detectable varies considerably among individuals, but except in extreme cases, no consistent relation between taste acuity alone and palatability judgments was indicated *(156)*.

The sense of taste may be relied upon to detect differences of concentration which represent but a small percentage of the threshold value *(284)*. The sense of taste was able to discriminate changes as low as 1 percent in concentration of sodium chloride solution ranging in concentration from 0.13 to 0.20 percent. With sodium chloride, sucrose, lactose, lactic acid, and quinine sulfate solutions, 10-percent changes in concentration were readily detected *(284)*.

Bitter solutions not only have a low threshold value but are slow of adaption as well, making it possible for a judge to distinguish between samples even when the concentration difference is small *(284)*. There appears to be little correlation between a person's ability to taste two kinds of bitter. Perception of sweet and bitter in mamose depends upon individual thresholds, which may be different for these two sensations *(26)*.

Known chemical substances, the concentration of which can be controlled, are better than flowers and natural odors for testing judges *(26)*. People differ greatly in respect to the threshold at which they can first detect the odorous substance *(26)*.

Reliability

Ten of the references reviewed reported tests for reliability of judges *(65, 69, 82, 103, 154, 182, 188, 193, 201, 227a)*. Various tests were used,

including the following:

(1) Ability of judges to recognize duplicates — a good judge will recognize 18 or more pairs out of 20 pairs *(103)*.

(2) Ability of judges to arrange samples in correct order of concentration of sweetness, sourness, etc. *(206)*.

(3) Analysis of scores on duplicate samples *(65, 182)*, and deviation from panel average *(65, 227a)*.

(4) Deviation between duplicate samples *(227a)*.

(5) Use of standard reference sample of predetermined score *(227a)*.

(6) Use of questionnaire to discover eccentricities of taste *(201, 206)*.

(7) Testing by period of training *(154, 188)*.

(8) Control-chart method *(193)*.

Another method of testing reliability of judges includes a sensitivity test confirmed by popular opinion on a series of samples. The number of times an individual's vote corresponds with popular vote is set up as a percentage of the number of samples tested. Candidates with percentages above 75 percent are chosen as panel members *(69, 201)*.

In answering questions on taste, using standard deviation is a more reliable measure of discriminating power than the statistic obtained using three categories of judgment such as "yes," "no," and "uncertain," in threshold tests by "constant stimuli" method *(202)*.

It is suggested that panel judges be calibrated by hundreds of examinations which are recorded in written form. A judge may be required to have an average deviation of not more than one point from the average of the group and a standard may be set for his ability to report judgments on duplicate samples. Good flavor memory is an important attribute *(152)*.

There is general agreement that the first requirement of a panel judge is reliability. A second important attribute is validity; judges should be able to produce scores close to an established standard *(193)*. A good judge should be able to detect odor-strength differences of perhaps 15 or 20 percent *(201)*. He should have few food prejudices *(188)*. Indications are that the judge who can score reliably may not be able to criticize accurately but that the judge who can criticize the samples with fair accuracy may be able to score reliably as well *(293)*.

Tea tasting is an art that can be acquired by anyone with a sensitive palate, but observation and experience are important *(230)*.

Results of selected judges on baked products were compared with results of the entire group of judges, and also ability of judges to duplicate findings on successive tests was determined *(66)*. Analysis of variance of judges' scores will give an accurate picture of reliability, consistency, and discriminating ability of each judge *(217)*. Chi-square was also used to test homogeneity of panel *(217)*. A few experiments indicated that the average technician could repeat his findings on replicate samples prepared during the same day *(67)*. This was also true of apple judges who were tested for reliability with the use of duplicate samples. A large majority of the laymen judges were consistent in their decisions *(17)*.

Judges were tested with wheat products about 20 times over a period of 4 months, then judgments were analyzed and persons giving 55 to 80 percent of correct judgments were chosen *(191)*.

Test for consistency of judges of dairy products increases the efficiency of the technique. Suitable subjects are those whose scores for same

samples are significantly correlated. Their agreement or disagreement with others, however, should be disregarded to avoid biasing results *(31)*. Two separate abilities are involved in taste testing of dairy products: (1) Recognition of quality and the placing of a numerical value on it, and (2) identification and description of items that make up that quality *(293)*.

Judges were rated excellent, good, and fair, depending upon ability to identify average concentration of primary tastes *(169)*. The judging characteristics of the individual may be investigated numerically by computing the correlation coefficients and regression equations relating their assessments to the average of those of all other members of the same panel *(156)*.

Size of panel required for specified accuracy

The number of judges serving on panels reported in the literature reviewed ranged from 3 to 50 *(5, 6, 65, 82, 89, 154, 161, 182, 188, 210)*. The majority of panels were made up of 4 to 12 members *(5, 6, 82, 89, 161, 182, 188, 210)*.

The permissible size of the panel depends upon the ability and training of the members and the minimum acceptable precision. Assuming a given degree of competence for individual panel members, the larger the panel, the more reliable the mean values obtained. If only three or four acceptable panel members are available, rescoring of each sample two or three times will give approximately the same reliability of final mean score as scoring once by a larger panel *(65)*. A small panel of selected tasters is more reliable than a larger group, some of whom are of doubtful ability *(227a)*. While the panel need not be large, its members should be experienced in the tests being made and any members whose judgments prove inconsistent should be eliminated *(154)*.

For paired judging, a high percentage may not be experienced tasters, and the larger the total number of participants, the more reliable will be the results *(152)*.

The type of product, number of qualified people, and amount of the product available for testing influence the procedure employed *(153)*. Panel methods are used in which 6 to 10 selected persons score the product. Products are also scored using paired and triple comparison procedures with larger groups *(153)*.

Variation with character of product and objective of study

Size of panel differed widely for wheat products, ranging from 16 to 96. Apparently use of a large number of judges does not increase the validity of results of tests on bread flavor *(167)*.

More reliable egg judging is reported with use of 5 good judges scoring twice than 10 poorer ones scoring once *(248)*. The number of members reported on panels for judging eggs ranged from 2 to 50.

Opinions differ as to number of judges required for reliable results when judging primary tastes. One report reviewed suggests that a large number is needed *(212)*, while another states that 12 can constitute a useful panel *(251)*, and still other tests suggest that reasonably satisfactory results were obtained with 15 to 20 tasters *(59)*.

Training of Panel Members

Training procedures in common use

The ability to taste may be developed. Training consists essentially in developing the senses accurately in order to know and recognize the various flavors that may occur *(289)*. A successful method of arousing individuals to test their acuity to define and identify flavor sensation has been to use solutions of salt, sugar, acid, and quinine for tasting and flavor essences for smelling *(289)*. One method of training judges is through the use of standard reference sample or samples of predetermined scores *(227a)*.

Amount and kind of training needed

A person can be trained to judge flavor by first examining samples in which a single flavor predominates. The person should recognize and estimate the intensity of his sensory reactions to this flavor. This flavor is then diluted and again the intensity of the reactions to the lessened stimuli are estimated. Next, the person should try to arrange in order of intensity, a series having easily recognized steps. As the training progresses, the steps should be narrowed. This must be done over and over again over a period of several days or even months and must be done with each particular flavor *(247)*.

Professional tasters are not persons of abnormal taste sense, but by long and careful training they have been able to develop their ability to distinguish one taste from another with great accuracy and to determine through dilution the continued persistence of a specific taste *(156)*. As far as inferior tasters of beer were concerned, no improvement took place; for the expert tasters, there appeared to be a tendency toward improvement *(151)*. In judging wheat products, the trained judges were no keener at discerning small differences in taste and odor than the untrained group, perhaps because differences were below the threshold at which the factor under test was noticeable (*167*).

With baked products, it is seldom feasible to train to detect specific off-odors or flavors, because they can scarcely be anticipated, no standards for comparison may exist, and time is required for training *(217)*.

Training for egg judging included training to primary tastes *(156)* or a practice period of about 1 week *(248)*. The necessity for preliminary training is shown in a study of scoring flavor of scrambled egg, in which the error was highest in the first week and lowest in the last part of the study *(98)*.

Oil judges were trained by appraising a group of oils that had been previously rated by experts *(146)*. Of the 40 people who took the flavor tests, 12 were selected as tentative panel members. This group was given training for a period of several weeks *(80)*.

Methods of Checking Performance of Panel Members

Deviations in scores

The performance of judges has been checked by deviation in scores on duplicate samples or deviation from mean panel scores *(65, 152, 156, 182, 227a)*. Scores of panel members can be averaged and their deviation

from the average determined. Fisher's statistical formula is applied when it is desirable to determine the significance of the difference between two mean scores *(152, 153)*. Scores that are completely off are eliminated from the average *(153)*.

The triple comparison procedure was found highly satisfactory; eight or nine of the tasters who gave the most consistent results in the triple comparison procedure gave identical results in the panel method of scoring the same product *(153)*.

In testing baked products, it was considered that low deviation from the judges' own mean scores indicates either high degree of reproducibility of judgments or lack of discrimination among different samples. If all scores are high, a low deviation could also mean that the judge is easily pleased. The fact that he could discriminate would not appear through an analysis of scores. Procedures for obtaining scores, rather than the analytical methods, were at fault in not avoiding this error. If mean scores are different and this is accompanied by low deviations from his own means, the judge is considered reliable in detecting differences *(217)*.

The following method of analysis for evaluating the ability of the judges has been used: Those scores showing poor checks on the duplicates were eliminated by subtracting the lowest score from the highest for each tester in each quality category, such as flavor and texture. Twenty percent of this difference was rounded to the nearest score point. This figure then represented the allowable variation that a tester could record for duplicate samples, and a table was made for use as a basis for accepting or rejecting a tester *(182)*.

Control chart

The control chart provides a method of measuring agreement of an individual with others. It is useful in the selection of a good tasting panel, determines when specific tasting scores must be examined, and minimizes losses from failure to pool results of "good" tasters and from pooling results that should not be pooled. It also helps to indicate the length of time needed for training *(193)*.

Correlation and regression coefficients

Correlation *(300)* and regression *(202)* coefficients are frequently used to check the performance of judges *(156)*. The ordinary method of testing a panel by the criterion of correlation in trials by duplication was considered questionable *(82)*.

Analysis of variance of individual scores

Analysis of variance has been applied to individual scores *(182, 229)*, used in egg studies *(156, 163, 180, 226, 273)*, and in preliminary sensitivity tests on fats and oils *(213)*.

Preparation of Samples

Size of samples

Dove states that samples for food testing must be large enough to reach all taste organs but not so large as to cause fatigue *(103)*. A sufficient amount of sample for two or three bites is usually a normal quantity *(188)*.

Beverages. An 8-gm. sample of coffee was placed in a cup of boiling water *(232)*.

Cereal products. The size of sample varies with the product and with different investigators of the same product. In one study two slices of bread were served to the judge *(167)*. In another study judges were permitted to taste as many pieces of bread as they desired *(66)*, while a third investigator reports that judges smell the fresh surface of a whole loaf of bread *(191)*. On experiments with cake a ¼-inch slice was cut from the center of the cake and served to the judge *(197)*.

Dairy products. Two investigators feel that just a small amount of ice cream, butter, and cheese is adequate for tasting *(31, 105)*. In judging milk, 8 to 12 ml. were served, but as much as desired was given to reassure the judge *(288)*.

Egg products. Size of the sample varied with the character of the product. An effort was made to provide samples of uniform size. However, size was conditioned by the viscosity of the egg and the judge's preference *(187)*. Several reported serving a small amount of cooked yolk *(122, 227, 253, 271)* and raw yolk *(122, 187)*. One served a whole egg *(248)*. Another served half of a hard-cooked egg cut lengthwise *(189)*.

A 3- by 1-inch slice of cake *(96)* sliced from the same position in each cake *(205)*, one-half a muffin *(96)*, or a custard cup of custard *(163)* were served to the judges.

Fats and oils. Five to ten milliliters of oil or shortening were served by a number of investigators *(80, 89, 213)*, 20 grams by others *(179)*.

Fruits. Oranges were cut transversely and from each half was cut a wedge-shaped piece for tasting. Each judge was advised to taste several pieces of one sample before rating it *(143, 144)*. Pieces or parts of halved peaches, apricots, and nectarines were served *(165)*.

Meats. Many reported serving one slice of meat to the judges *(9, 23, 24, 68, 79, 106, 157, 186, 220)*. Others reported slices or cubes of definite dimensions: Steaks 0.6-inch thick (*41*), ¼ by ½ by ½ inch (*75*), 1 by 2 inches (*70*), ½ of 1 inch square (*70*), slice 5- to 7-mm. thick (*9*), slice about 3/32-inch thick (*252*), one slice 0.6-cm. thick (*216*). Three reported servings small enough to be eaten in one mouthful *(75, 77, 195)*.

Poultry. Small cross sections about ¼-inch thick were served *(107, 286)*.

Vegetables. Approximately 100 gm. of raw fresh greens *(291)*, 1-inch lengths of broccoli stems *(15)* and 1-inch sections of ears of sweet corn *(101)* were tested. When tasting peas the tasters were requested to place several peas from each sample in their mouths at the same time in order to get a representative sample *(30)*. The size of sample from roots and tubers varied from small bites of potatoes *(134)* and one-sixth of a slice of carrot about ¼-inch thick taken about 1 inch from top of root *(72)*, to half of a potato *(255, 299)*.

Primary tastes. Many investigators reported the use of 5 to 10 cc. of solution for taste tests *(2, 59, 84, 95, 112, 166, 169, 237, 241, 284, 297)*. Others reported the use of 0.6 to 2 cc. *(25, 241, 243)*, about one-half of 0.2-gm. tablet of thiouracil, 1 teaspoon 0.005-percent solution of phenylthiocarbamide, and 20 cc. of more dilute acids *(21)*. Still another suggests

using as much as is needed for matching test *(112)*. Trout and Sharp *(284)* point out that as the difficulty of placement within the series increased, the amount of solution required in arriving at final placement also increased. Therefore, limiting the taster to a fixed minimum of solution would seem to be a handicap in placing a series with the highest possible correlation.

Temperature of samples

Many investigators recommend serving the samples at a uniform temperature at which the specific food is normally served *(69, 103, 152, 154, 188, 201, 210, 227a, 295)*. Others feel that most materials are best tasted at body temperature *(85, 151)*. The gustatory nerves cease to function at 50° C., and taste is also strongly reduced below 15° *(151)*.

Beverages. Freshness in coffee is more readily detected when the coffee is hot, while staleness is more easily noted when the beverage has been cooled *(232)*. On the other hand, extreme refrigeration can be shown to rob the consumer of the ability to distinguish flavors, and it is probably true that fruit squashes, flavored mineral waters, wines lending themselves to icing, lager and nondeposit beers exhibit their truest flavor character, and are most refreshing at from 50° to 55° F. *(20)*.

Cereal products. Crocker *(85)* reported that cakes should not be scored when warm. Bread was sampled at room temperature *(167)*.

Dairy products. Downs feels that cold materials should be warmed to a temperature approaching 98° F. Ice cream should be held on the tongue until it is warmed to body temperature, and cream should be at a temperature of 70° to 90° F. for tasting *(105)*. Ice cream was served at a temperature high enough to permit ease in dipping, but not soft (*105*), and in an ice bath (*96*). In one laboratory ice cream was served at approximately 15° F. (*178*). Butter samples were served at approximately 50° F. (*129*). Cheese samples were served warm (*105*). Milk was served at room temperature, 21° C., in one instance (*284*), and at 90° F. in another (*85*). Josephson (*164*), in working with dry milk powders served at 5°, 15°, 22°, 37°, and 45° C., concluded that judgments of milk powder taste qualities are less critical at 22°.

Egg products. Temperatures of serving varied with the product. Cakes were stored overnight before judging *(205)*; rolls and muffins were served hot *(96)*. Mayonnaise was allowed to stand 4 hours at room temperature *(96)*; custards were allowed to stand 5 hours before testing *(73)*, cooled *(96, 163)*, or stored overnight in a refrigerator *(11)*. Scrambled eggs *(3, 4)* and cooked yolks *(227, 253, 271)* were tasted while warm; poached eggs were allowed to cool slightly before judging *(296)*.

Fats and oils. Harding reports that all oil samples should be warmed to the same temperature, as correct temperature renders flavor and odor more detectable *(146)*. Soybean oil was served at 80° C. *(213)*, at 45° to 50° C. *(80)*; salad oil was served at 45° C. *(89)*; samples of linseed oil were judged hot and again when cooled *(179)*.

Fruits. Frozen peaches, apricots, and nectarines were served cool *(165)*. Dried apricots were served warm *(258)*.

Meats. Hot samples of meat have been used by a number of persons

(24, 70, 157, 186). By others, samples were served warm *(79, 158),* at room temperature *(220),* or cold *(157).*

Poultry. Time at which birds are inspected after cooking is important, but opinions differ slightly on the optimum time for tasting. If samples are inspected immediately after removal from the oven, aroma is pronounced and may reflect quality more vividly than flavor does *(278).* Vail and Conrad suggest that odor should be evaluated after the bird has cooled slightly *(286).* Several reported palatability tests of hot samples *(107, 150, 192, 214),* and one served warm samples *(263),* but all samples should be at the same comparative temperature *(184).*

Vegetables. All samples of sweet corn were allowed to cool to room temperature for tasting *(49).* Broccoli, asparagus, snap beans *(15),* and peas *(30)* were served warm. Wright and coworkers *(299)* in their study of potatoes served all samples hot, while Smith, Nash, and Dittman *(255)* allowed their samples of boiled potatoes to stand in the air for ½ hour.

Primary tastes. The temperature of the solution is an important factor *(111, 212).* Several investigators recommend that solutions of primary tastes should be at room temperature for testing *(7, 85, 100).* But other investigators report that the optimum temperature for tasting seems to vary somewhat with the solution: 21° C. for salt solutions, lactic acid, and quinine; 35° for sucrose and lactose solution *(284).* In contrast, Salmon and Blakeslee report that changes in temperature of solutions appeared to have no material influence on thresholds. Reaction to cold solutions was delayed in some cases until the solution was raised to mouth temperature (*241*). The majority of panel members could detect bitterness in a cold, saturated solution, while the rest could detect it in a hot, saturated solution with crystals in suspension or in a similar hot solution in weak alcohol (*25*).

Method of cooking or other preparation of samples

In a system designed to measure the specific odor, taste, and visual properties of foods and beverages and also the preferences of the typical consumer, the equipment controls the samples at any temperature desired, contains a bank of 48 flasks which dispense exact amounts, has an enunciator which records the judgment of the observer, and is constructed so that the technician and observer are completely separated. Timing is controlled by use of a turntable device. Other possible sources of variation in procedure, as glassware, color, and rinsings, are held constant *(300).*

Strong-tasting ingredients and their standards are preferably diluted with an inert solid or dissolved in water to bring them down to a comfortable concentration for comparison *(85).*

Beverages. Freshly boiling water should be used in the making of tea for testing. The tea should be infused for 6 minutes and then poured off the leaves *(230).* In the preparation of coffee, samples of the stored coffee that are to be compared with the fresh roast are weighed into separate cups; boiling water is poured into all. In preparing standard samples, if the coffee to be tested is obviously quite fresh, 8.0 grams of the fresh roast is put in the first cup and the amount decreased by 0.8 gram in each succeeding cup. If the sample to be tested is quite stale, 4.0 grams of the fresh roast is put in the first cup and a series decreasing successively by 0.8 gram per cup is made (*232*). If the sample exhibits "stale"

flavor, another series of test cups is made up. In each of these cups is first put the amount of freshly roasted coffee found in the first series to match most closely the flavor of the sample, for instance, 3.2 grams for a coffee of 40-percent freshness. Then successively increasing amounts of a thoroughly stale sample of the same blend are added *(232)*.

Cereal products. All possible variables in the preparation of pastry for testing were controlled so that differences were due only to the different fats used *(217)*. In preparing cakes, all procedures were standardized by weighing ingredients, counting mixing strokes, and controlling time and temperature of baking *(47, 197, 203, 204)*. Bread samples used for toast were prepared under controlled laboratory conditions *(67)* as were macaroni samples *(149, 218)*. To eliminate the effects of oven position in the baking of bread, each row of loaves across the oven contained loaves from each dough and the relative order varied from row to row. Each row was numbered and the loaves from any one row were judged together *(191)*.

In the baking of bread, the end point of cooking was determined by amount of time in a regulated oven *(66, 167, 191, 218)*. Maiden *(191)* suggests baking the bread in closed tins.

Cookies were prepared by adding different flavors to a basic formula. After being cooled for the same length of time after baking, they were stored in cans having tight covers, then examined the next day *(125)*.

Dairy products. All fresh milk samples were cooled after being drawn and stored in a 40° F. refrigerator until judged (*74, 105, 288*). Dry milk powder was reconstituted with distilled water and chilled *(224, 225)*. Pearce and others *(226)* reconstituted dry milk and dried eggs as a milk shake mix, sweetened but unflavored, for judging. Ice cream was prepared using different concentrations of cane sugar and compared for sweetness *(178)*.

Egg products. Egg samples were prepared and tested under standardized conditions *(163, 205)*; uniformity in quality and quantity of ingredients and such factors as temperature of ingredients, methods of mixing, and time and temperature of baking, were controlled as far as possible *(11)*.

Eggs for boiling were immersed in boiling water in a wire basket or mesh cloth and cooked for a definite length of time *(122, 189, 227, 248, 253, 271)*.

Fresh eggs were poached according to standard procedure. Those to be scored for appearance were poached with salt and those for taste were poached without salt *(189)*. Reconstituted dried whole egg was poached in a steam egg poacher *(259)*.

Baked custards made with reconstituted dried egg powder were placed in custard cups in a pan of hot water and arranged systematically in the oven *(11, 73, 98, 99, 163)*.

Dried eggs were generally made into scrambled eggs for quality testing *(3, 18, 19, 35, 36, 39, 96, 98, 99, 116, 156, 181, 194, 209, 262, 273)*. The eggs were put in beakers which were placed in gently boiling water and were stirred until coagulated (*3, 4, 19, 99, 262, 273*). Other workers (*96*) prepared part of their samples in a fry pan, with and without bacon fat.

In cakes, dried egg powder was substituted in regular cake formulas *(11, 96, 205, 267)*. Muffins, yeast rolls, and popovers were also made by standard procedure using dried egg shell *(96)*.

End point of cooking was determined by time for boiled and poached eggs *(205, 248, 253, 296)*. Baked custards were determined done by internal temperature *(11, 73, 141, 163)*. For scrambled eggs, the end point of cooking was determined by tenderness, that is, consistency of typical scrambled eggs *(3, 99, 209, 262)* or serving consistency *(262)*. There is no precise method of testing the end point of cooking scrambled eggs *(98)*.

Fats and oils. Hot and cold pressed raw linseed oil, prepared according to standardized processing procedure, was heated in the oven for serving *(179)*. The end point of cooking for baked and fried products containing experimental fat and lard was determined by time *(13, 168)*.

Meats. Controlled standard methods were employed throughout the preparation of the meat samples *(14, 24, 68, 70, 75, 76, 106, 138, 157, 195, 215, 216, 233, 252)*. Many investigators determined the end point of cooking by the internal temperature of the meat as recorded by a thermometer *(9, 23, 24, 41, 68, 70, 75, 76, 78, 79, 106, 127, 138, 147, 157, 195, 215, 216, 219, 220, 234, 242, 252)*, while others used time in determining the end point of cooking *(9, 61, 87, 233, 234, 252, 294)*. Others recommend that bacon for judging should be cooked until crisp and an even light brown. It should not be translucent *(9)*.

Poultry. In the preparation of poultry for taste tests, several investigators determined the end point of cooking by time *(12, 90, 117, 200, 286)*; others used a thermometer to determine internal temperature *(107, 142, 171, 183, 185, 192, 244, 264, 265, 286, 295)*. Harshaw and others *(150)* determined the end point of cooking by tenderness, that is, when the wing joints had softened and the thigh meat could be pierced with a skewer. No salt or other seasoning was added during cooking *(90, 107, 150, 214)*.

Vegetables. Methods of preparation of all vegetables were as rigidly standardized as possible to insure identical treatment of all samples *(44, 56, 57, 60, 115, 134, 255, 272, 275, 291)*. Precautions were taken to use the same size and kind of cooking container, similar amounts of water and approximately the same length of time for cooking, and uniform gas flames *(255)*. Several investigators recommend developing uniform cooking procedures by preliminary tests *(15, 30, 299)*. No addition of salt or other seasoning was made prior to testing *(33, 51, 52, 56, 91, 101, 135, 266, 290, 298)*. In handling fresh greens Whitacre and coworkers *(291)* recommend that the same person prepare the greens for cooking to minimize variation due to personal factors.

End point of cooking was determined by time for root vegetables and greens *(15, 30, 32, 33, 43, 49, 50, 51, 52, 53, 55, 56, 91, 101, 114, 115, 127, 133, 172, 177, 231, 254, 276, 277, 285, 287, 298, 299)*, by internal temperature of baked vegetables as recorded by a thermometer *(57, 135, 261, 299)*, and by tenderness, that is, when baked potatoes felt soft to slight pressure of hand *(71)*, when boiled potatoes seemed soft by thrusting a paring knife into them *(71)*, or testing with a fork *(43)* or cake tester *(134)*. Whitacre and others *(291)* determined length of cooking period by piercing turnip greens with a knife or fork or by testing with the teeth.

All dehydrated vegetables such as sweetpotatoes, white potatoes, beets, and cabbage were reconstituted and then prepared by standard methods *(51, 53, 55, 56, 57, 114, 115, 266)*. Steamed, baked, and boiled potatoes were cut in half; one half of each potato was judged for color and texture

and the remaining halves were peeled, riced together, and the composite was judged *(299)*.

Primary tastes. Solutions of sucrose and of sugars to be tested were made fresh just before judging *(93, 236)*. In one laboratory all glucose solutions were prepared 16 hours before use and were therefore equilibrium mixtures of alpha- and beta-dextroglucose; also all lactose solutions were so prepared. Solutions were generally used within 24 hours and never when more than 2 days old *(58)*. Solutions of phenylthiocarbamide were made up from a stock solution of 1: 5,000. Solutions in the small bottles prepared for the judges were frequently renewed, since they tended to lose their strength, perhaps on account of soluble substances on the soda fountain straws used in the test *(25)*.

Various kinds of water have been used in making up solutions for taste tests: Distilled water for sucrose solution *(236)* and for sucrose, lactose, and lactic acid solutions *(284)*; Poland water for the major part of the chloride studies and for the quinine-sulfate studies *(284)*; and artesian well water for phenylthiocarbamide solutions *(25)*. Trout and Sharp *(284)* warn that not all good drinking water is necessarily adapted for use in taste studies. They suggest that suitable water for taste studies may be obtained from local springs. Trout and Sharp further indicate that the sense of taste may be relied upon to differentiate between the various tastes of water.

Concentrations of the solutions in a series were varied according to two procedures: (1) Geometric progression, obtained by dividing each successive concentration by the factor selected; (2) arithmetic progression, obtained by making equal increments in change in concentration *(284)*. In all cases a factor of two was used in making dilutions *(29)*. Samples of egg were prepared by adding one of the primary taste test substances in an amount approximately equal to or definitely above the median threshold in the preceding trial *(156)*. Solutions were half as strong as threshold dilution determined by a previous test *(241)*.

Sugars were made up into solutions by even percentages from 1 to 8 percent *(84)*. Solutions of sucrose were made in 5-, 10-, 15-, 20-, 25-, 30-, and 40-percent concentrations as standards *(93)*. To eliminate the subjects' having to drink large amounts of water, concentrations were increased in large steps, skipping one or two solutions each time, until the point was reached where the subject recognized the difference between the two samples. The concentration was decreased in small steps until the lowest concentration at which subjects could still tell the difference was determined *(236)*. In another study each solution tested was four times as strong as that previously tested *(25)*.

Serving of samples

All food for testing should be served in the form in which it is customarily eaten, at normal temperature and strength (at or near threshold concentration) in a neutral medium *(85)* because at different concentrations some substances may give qualitatively different sensations, for example, saccharin *(202)*. Hallmark Testing Service *(188)* recommends serving the samples of food first plain, and then in form as usually consumed.

Beverages. Tea is served hot and tasting is done by "slurping" a large spoonful which is held in the mouth for only a few seconds and then expelled. Hot coffee is sipped to evaluate taste factors and must be

drawn back into the mouth slowly. Both coffee and tea should be judged for aroma first *(85)*.

In beer judging, only beers with a suitable degree of difference between them should be served *(151)*.

Cereal products. A system of judging bread is recommended by Maiden *(191)*: The bread should be judged when about 18 hours old and representative loaves should be used. No more than four people judge any one loaf for smell and taste, so that each person has a reasonable-sized piece to smell, and before a person passes an opinion of the smell of a loaf, a slice about 1 inch thick is cut off. Thus, each person has a fresh surface to smell and one that contains a fair amount of the natural gas of the loaf. Each observer judges two sets.

Two investigators recommended serving slices of fresh bread to the judges. Ingels, Irwin, and Landis *(159)* found that the best procedure was to wrap two slices of bread together in waterproofed cellophane paper, held together by a rubber band. Cathcart *(66)* served sliced samples in sterile, covered glass jars.

In the serving of cake, individual pieces were wrapped and submitted to the judges *(204)*.

Dairy products. Milk samples should be well mixed before being opened and part of the contents poured into a glass. Odor should be observed first, then flavor tested by sipping slowly, allowing the milk to remain in contact with the tongue for a short time *(105, 284)*. Weaver recommends inverting the bottle of milk, then removing the cap and pouring the sample into a beaker *(288)*. Cream is tasted by dipping a glass or composition rod into the sample and transferring a small amount to the mouth *(105)*.

Firmness of cheese was judged in three ways by Scott Blair, Coppen, and Dearden *(246)*: (1) By plunging a skewer of standard pattern into the cheese, (2) by pressing the top of the cheese with the thumbs and fingers, and (3) by having the subjects handle borings of the cheese as they pleased.

Eggs. For serving the baked custard of any one baking, a judge was given the three custards from a given horizontal row as the cups were placed in the oven *(163)*.

Boiled eggs were placed in egg cups for serving; the shell of the blunt end was removed below the air cell and the warm yolk was mixed with a spoon or a glass rod *(227, 248, 253, 271)*.

For serving, raw egg yolks were placed in small glasses and covered immediately. Glasses remained covered except when sample was being removed. Egg white was discarded. Before testing, each yolk was mixed to insure uniform consistency. Glass rods were used to convey the sample to the mouth (*122, 187*).

Fats and oils. Samples of oils are served in beakers which are placed in electrically heated aluminum blocks. Each member of the panel has his own tray which may be removed from the source of heat and the tray will then hold the heat during the tasting period *(80, 213)*.

Fruits. Oranges were cut transversely and from each half was cut a wedge-shaped piece for tasting *(143, 144)*.

Meats. The slicing of samples for serving for paired judging is very important. All investigators report that the samples for each judge must

come from the same relative position of the meat each time. Thus, samples were paired not only for position of the slice in the roast but for position of the sample within the slice *(41, 68, 70, 75, 76, 77, 78, 79, 106, 147, 158, 195, 216, 220, 294)*. No salt or seasoning was added *(23, 24, 70, 138)*. In carving, care was taken that the sample to be judged was not contaminated by the developed flavor from the browned outer surface or extra fat *(70, 158)*.

Miscellaneous foods. For reheating and serving of frozen precooked meals, the foods were unwrapped and placed on glass pie plates. All plates were covered partially with aluminum foil. Solidly frozen foods were reheated 25 minutes in a 375° F. oven; others were placed in a preheated Maxson Whirlwind oven and heated for 15 minutes at 300° *(126)*.

Poultry. It was generally agreed that the carving method used throughout the judging period should be standardized so that at each judging session the judge will receive a slice from the same relative position of the bird for scoring *(90, 107, 150, 192, 200, 264, 265, 286, 295)*.

Vegetables. To prepare for sampling, buds and stems of broccoli were placed in separate warmed bowls, the buds were chopped with a special chopper while the stems were cut into 1-inch lengths *(15)*.

Potatoes were served in numerous ways for judging. Boiled potatoes were placed on a plate, mashed with a fork, and covered *(134, 261)*; passed through a ricer (*55, 56*); or broken up and well mixed with a wire potato masher *(43)*. Baked potatoes were cut into small pieces with scissors and served in heated glass dishes *(43)*. French-fried potatoes were drained and served in a hot glass casserole *(43)*.

Sweetman *(270)* reports that the texture appearance of potatoes has been judged from (a) cut surface of the cooked tuber in cross section, (b) scraped surface, (c) crushed mass, (d) mass disintegrated with a fork, (e) mass put through a ricer. The relative smoothness or granularity of "feel" on the tongue when rubbed against the roof of the mouth can also be tested.

Primary tastes. When conducting taste tests one investigator recommends that only one series of acid should be served at one time and sour samples should be served last *(112)*.

Conditions of Judging and Judging Room

Time of day

Opinions vary as to the best time of day for judging. Investigators recommend an early morning judging session *(1, 201, 241, 288, 294)*, from midmorning to late morning *(13, 80, 97, 99, 114, 126, 154, 157, 167, 182, 227a, 236, 241, 248, 259)* and midafternoon to late afternoon (*49, 80, 114, 154, 182, 227a, 236, 241, 259*).

Others suggest that judging should be held 2 to 3 hours after eating *(95, 166, 297)* and 45 minutes to 1 hour after eating *(175, 236, 288)*. In the judging of fats or oils, preferred times were not less than 2 hours after a meal and 1 hour after eating or chewing gum *(80)*.

Other recommendations include: Time when food would normally be eaten, for example, hot cakes in the morning *(188)*, when subjects feel their best *(22)*, not when extremely hungry *(188)*, and for pastry judging whenever convenient for the judges *(217)*.

Judging of primary tastes was held during a dinner between courses *(26)*, while fruit samples were served as part of a dinner *(258)*.

Utensils used

Container and eating utensils used for sampling should be completely tasteless and odorless *(22, 188)*. The use of gray, white, or glass containers, preferably on gray background, is suggested *(188)*. Containers must be clean and attractive *(154)*. Food products should be served in suitable containers, with due consideration to individual properties of each product, such as sensitivity to air, light, and dehydration *(22)*.

In the judging of dairy products, samples were presented in small containers *(31)* such as a cup or glass *(105)* or 50-ml. beakers *(288)*. Fat and oil samples were served in 5-ml. beakers *(89)* and in 100-ml. beakers *(179)*. Primary taste solutions were served in 50-ml. beakers on individual trays with glass or large beaker for distilled water and paper cups for discarding solution *(95)*. Egg samples were presented in cups with small paper spoons for tasting *(248)* and with tin spoons, which were washed after each taste *(248, 253, 271)*.

Warm plates were used for meat samples *(70, 157, 252)* and vegetable samples *(15, 52)*. Legumes were served in an open saucer and the judge used a fork or finger to separate hard from soft parts *(30)*. Baked custard was served in the cup in which it had been baked *(163)* or turned out on paper plates *(62)*. Raw egg samples were served in covered dishes *(122, 187)* or placed in egg cups and tasted with small spoons *(227)*. Freshly baked bread samples were served to the judges in glassine bags with the tops folded down and fastened with clips *(167)*. Turkey samples were placed on a toothpick for judging *(214)*.

Tests for the primary tastes were made with soda fountain straws *(25)*. A glass rod was used to taste cream *(105)*, turkey drippings *(192, 214)*, and raw egg samples *(253, 271)*. The rods were cleaned in warm water after each sampling *(253, 271)*.

Coding of samples

Code numbers should not be suggestive to the panel *(69, 188)*. The first of any series suggests first choice to many minds, so that it is inadvisable from a psychological standpoint, to use a, b or 1, 2 or any other logical series of markings *(188)*.

In judging pastry, papers with numbers were placed beneath wire racks containing samples *(217)*. Symbols were used in a consumer preference test on eggs *(18)*. In another egg test, each judge was blindfolded so that he could not know or be influenced by appearance *(187, 209)*. Some tests for primary tastes have used samples marked with code numbers visible to the judges *(58, 59, 84, 166, 169, 175, 293)*, and others have used hidden markings *(21, 237)*, such as adhesive tape covered with paraffin on bottom of beaker *(95, 284, 297)*. In some of the tests, the judges knew what they were testing for but did not know the order *(95, 166)*.

Time after smoking

A time lapse of 1 hour or more is suggested after smoking *(85, 212)* and another reference recommends an interval of 2 hours *(188)*. The opinion is given that heavy smokers do not make good tasters *(69)* as smoking blunts delicacy of taste and the effect varies with individuals *(59)*. There

is no strong correlation between the use of tobacco and sensitivity to phenylthiocarbamide *(241)*.

Discussions at judging session

It is generally agreed that all talking during judging should be prohibited *(5, 9, 22, 69, 80, 124, 126, 140, 182, 184, 188, 192, 276, 286)* but that discussions are permissible after the judging session *(69, 89, 101, 146, 154, 169, 188, 213, 214, 215, 248, 288, 296)*. This plan helps maintain interest and enthusiasm of the tasters *(188)*, favors discovery of important values and develops better interpretation of observation among the judges *(89)*, helps to make panel members aware of points to be judged and leads to uniformity *(275)*.

In the judging of meat, the judges were permitted to talk freely but could not identify their samples with neighbors' because of coding *(77)*. Discussions were held during judging of root and tuberous vegetables, but judges were asked to record unbiased opinion *(43)*. Discussions also were held during the judging of primary tastes *(26)*.

Conversation was limited during the judging session of eggs, but after the numerical score was given, the judges were asked to describe flavor *(187)*.

Time allowed for tasting

Recovery from a tasting is a matter of minutes and a rhythm method of tasting should be used. One or two tastes each minute are suggested or even one each 5 minutes for very strong tastes or for thick or clinging materials *(85, 148)*. Solid substances must be thoroughly masticated *(22)*. Sensations must be successive (not simultaneous) so that there is a real sensation of difference aroused by the shock of transition from one perception to another which is unlike the first *(201)*. The taster should act on his first sensation of an odor, not allowing himself to become accustomed to it *(201)*. If odor is to be judged, food should be sniffed before it is tasted. All samples should be smelled in turn and tasted only after smell impressions have been recorded *(22, 188)*. Liquids should be retained in the mouth longer than usual *(22)*.

As much time as desired is allowed for the tasting of milk samples *(288)*. The actual time required by an experienced judge to ascertain the taste of a sample is surprisingly short. Such a judge has come to realize that the taste-reaction time not only varies with different tastes but also that it is fixed within certain ranges for specific tastes. Sometimes the experienced judge finds it necessary to retaste a sample, but often judgment is passed on the basis of the one taste reaction. The inexperienced judge tends to waste the precious first moments of tasting. Meanwhile, the onset of taste adaption may make it more difficult for the beginner to diagnose correctly the often delicate first taste sensation of the product *(279)*.

In milk tasting, generally only one tasting was necessary to arrive at a judgment, but in several instances when the taste reaction was not pronounced, a second or third taste was made. Tasting was done continuously from the start with only occasional minor interruptions. There were no rest periods following tasting of each sample or after tasting a group of samples. The average time to make flavor judgments ranged from a low of 2.9 seconds for "salty" milk to 8.2 seconds for "excellent" milk. Milk having off-flavors of slight intensity required considerably

longer judgment time than did those having pronouncedly intense off-flavors *(279)*.

Judging of primary tastes was held every 15 minutes after meals, and at least every hour during remainder of day, with no eating between meals. No correlation was found between time of making judgments and the accuracy of judgments *(241)*. To taste and judge a single sample of a primary taste required usually between 5 and 10 seconds *(284)*. The amount of retasting necessary before arriving at a final judgment of a primary taste depended on the concentration, its range, and the number of samples within the series *(284)*. In judging fats and oils, the samples were held in the mouth for about a half minute *(213)* or for 10 to 30 seconds *(80)*. Ten minutes were allowed for judging meats *(9)*; only a few minutes were allowed for evaluating the odor of poultry *(286)*.

The time interval between samples is very important. For example, the taste qualities of normal milk powder have a tendency to "block" the taste mechanism so that a "carry-over" or "build-up" of the taste qualities of the previous sample results *(164)*. In tasting primary taste solutions an interval of 2 to 5 minutes *(112, 166, 212)* and 15 to 20 minutes *(59, 241, 243)* between samples is suggested. In the sampling of beer, it is not advisable to serve two sets of three samples in close succession as the taste is somewhat dulled after the first set *(151)*. Tests should be made as quickly as possible but unhurriedly to avoid fatigue *(251)*. Judges of egg products were urged to rest if they felt their taste buds tiring *(122)*.

Method of removing flavors from mouth

It is generally considered better for judges to eject samples rather than swallow them *(80, 85, 89, 122, 210, 213, 229)*. Tepid clear water is recommended for rinsing the mouth between samples *(182, 188, 210, 229)* and has been used in testing such products as bread *(66, 167)*, dairy products *(105)*, egg products *(122, 187, 227, 248)*, fats and oils *(80, 89, 213)*, meats *(9, 157, 219)*, poultry *(286)*, and vegetables *(114)*. In one study of dairy products a tepid water rinse was used only when an especially pronounced nauseating flavor was experienced *(279)*. Wiping the tongue with a paper napkin followed by a water rinse was suggested for egg products *(122, 187)*.

Some of the references reviewed suggest that the taster should be supplied with water as well as various types of "throat and mouth clearers." When testing fatty foods such as mayonnaise, the taster is given celery or raw apple slices; for beer, use olives; for fish, use dill pickles, but no mouth washes *(188)*. Other substances for removing flavors were reported, such as white bread *(22)* for meats *(9)*, poultry *(214)*, and eggs *(122, 227)*; crackers *(182, 188)* for poultry *(286)*; and slices of tart apple for meats *(9, 157)*, dairy products *(105)*, and poultry *(90)*.

Another opinion is that something may be gained by the use of artifices between samples, especially rinsing the mouth with water, to remove adherent material, but the best general technique is to allow the saliva to lave the taste buds in a natural manner *(85)*.

The partial retention of the solution being tasted may be sufficient to account for some of the discrepancies in reporting taste sensations. Particularly may this be true after tasting a poor-flavor sample without having previously rinsed the mouth *(284)*. When tasting solutions containing 1 percent or more of NaCl, from three to four rinsings were

necessary to free the mouth of the retained salt *(284)*. Judges did not swallow samples *(25, 93, 95, 112, 166, 169)*. A water rinse was used after each solution and between solutions *(21, 25, 59, 84, 93, 95, 112, 166, 169, 241, 243, 297)*. The tongue was dried with blotting paper *(241)*. Rinsing the mouth is not considered very helpful in the testing done by some workers in the field *(212)*.

Location of judging room

All food tasting should be held in a special room suited to the purpose of palatability testing *(5, 65, 75, 80, 103, 146, 152, 154, 188, 207, 213)* although investigators have reported that it is sometimes necessary to judge in the laboratory where the food was prepared *(43, 122, 207, 288, 291)* or in the dining room *(26)*. Testing for bread was held where judge worked *(159)* or at a convenient point in the building near a water fountain *(66)*. Showalter reported that only one taster at a time was allowed to enter the laboratory for a test *(251)*.

Seating arrangement

Many investigators report that individual tables or booths are desirable for independent judging *(5, 43, 65, 103, 152, 154, 167, 188, 217, 298)*. Sometimes all judges are seated at one large table *(26, 77, 237)*.

Provisions for ventilation, lighting, and temperature control

The laboratory should be entirely air-conditioned or well ventilated and free from all odors that might prevent accurate detection of difference in samples *(80, 89, 103, 105, 124, 146, 152, 188, 213, 227a, 229, 288)*. Palatability rooms should be maintained at moderate temperature for accurate judging *22, 151*). Crocker recommends moist air when detecting odor of samples (*83*). Controlled temperature is very important in the testing of butter. The butter should remain firm but not too cold when fine odors are to be distinguished (*105*).

Many investigators agree that optimum, controlled lighting is desirable, as brilliant lights are detrimental to best results *(65, 69, 89, 124, 154, 201)*. When color is a factor, MacBeth Daylight Lamps are best to use *(227a)*. Daylight lighting is recommended for dairy products *(105)*, good north light for tea tasting *(230)*. When fluorescent lights are used, the combination of two daylight lamps and one pink lamp will give the best results *(227a)*. In some cases, scoring is done in a darkened room to cover up differences in color of the samples (*152*). One author recommends for best results that the judging booths should be equipped with a spot lighting system with three degrees of natural light and two degrees of colored light, plus control of intensity (*103*). Others attempted to have the judging room light and pleasant (*286*).

Other provisions for judging

Judging room should be free from outside distraction *(22, 69, 85, 89, 124, 151, 152, 188, 201, 251, 284, 288, 297)*. Arrangement of judging room should minimize distraction and discourage tendency of judges to make audible remarks or otherwise convey impressions. Quietness, smoothness of presentation, orderliness, and regularity contribute to more accurate evaluation *(213)*. In work on eggs, only two people

other than the judges were allowed in the judging room, one who was serving the sample and one who was recording scores *(187)*. There should be no unnecessary conversation *(251)* and the judges should not be able to observe progress of the others *(297)*. Those about to be tested should not be allowed to see the test given to others *(25)*. Samples should be presented under natural conditions *(22, 85, 124, 152)*. When delicate flavor differences are to be discriminated in samples of different colors, blindfolding the tasters is desirable *(69, 201, 206)*. Paper cups should be provided for discarding samples *(227a)*.

Dove *(103)* recommends a neutral gray color for the walls and judging table tops so no color is added to the food.

Judges must follow strict rules of conduct. Methods that have already been worked out respecting smoking, drinking water, etc., must be rigidly followed *(159)*.

Summary of Factors Determining Accuracy of Tests

Number and kind of characteristics evaluated

Of the various factors to be studied in the organoleptic evaluation of a food, the following are of importance: Feel, texture, color, aroma, aftertaste, and the stimulating factors of pungency and heat *(63)*.

One author recommends the evaluation of either general quality or a maximum of two specific characteristics *(182)*, while in cereal products from two to six characteristics have been evaluated *(38, 47, 66, 67, 159, 196, 197, 217)*. Frozen foods at Cornell were evaluated by odor, flavor, texture, surface appearance, color, and general acceptability *(126, 127)*. In other studies on vegetables three or more of the following characteristics are usually included: Flavor, odor, texture, surface appearance, color, size and form, and general acceptability *(30, 50, 97, 127)*.

Uniformity of material, quality of food

Many authors recommend that food products be of uniform material and quality. Egg products were sweetened and to prevent differences in sweetness affecting tasters' judgment, all samples were made to a level of 7-percent sweetness in terms of sucrose *(226)*. To avoid or standardize the effects of cooking, enzymes, or other factors on meat flavor, care must be taken to include samples of meat that are as similar as possible *(158)*.

Precautions must be taken to insure uniformity of vegetables *(276)*. Kale was always picked in the early morning and, insofar as possible, the leaves chosen were of comparable maturity for each variety and all pickings *(133)*. In the case of turnip greens, the same person selected and trimmed the greens, minimizing variation owing to personal factor in handling *(291)*. Peas were sorted to remove the immature and overmature ones, washed, and thoroughly mixed *(15)*. Snap beans were of a single variety of known origin and growing conditions and were graded for quality differences *(97)*.

In scoring tests on vegetables, it was shown that salt had no effect on average values given for color, shape, and odor, but there was a trend for higher ratings for flavor, texture, and acceptability in salted samples. There was no evidence that either addition or omission of salt enabled judges to detect small differences better in one case than another *(290)*.

Standardization of terminology used to describe quality

The British scoring system has been widely used with egg products *(4, 259)*.

Five quality groups (excellent, good, fair, poor, bad) with 18 descriptive flavor defects have been used in grading milk. If the descriptive term did not properly describe the defect, the judge used his own term or said "unidentified" *(288)*. In another study, standards for "strong to very strong," for "distinct to pronounced," and for "slight" oxidized flavor in milk were suggested. Taste judgments of oxidized flavors designated as "distinct" and as "strong" were found to be more accurate than judgments of oxidized flavor designated as "slight" or as "doubtful" *(284)*.

Number of samples, number of replications

Many investigators feel that the number of samples at one judging session should be limited: Two samples *(1)*, four or less *(22, 69, 182, 188)*, two to four and not more than six *(207)*. Crocker *(85)* states that eight specimens in a series are as many as may be worked effectively. Another author *(227a)* concludes that only as many samples should be served as the judges can taste without becoming fatigued.

It is not possible to duplicate taste tests many times because the senses become rather rapidly dulled and wrong impressions would be obtained if one attempted too many times to taste the samples *(1)*. Time and accuracy favor working with a small number of samples, always checking them against each other *(85)*. The stronger the taste and odor of a substance, the smaller the number of samples an individual can taste before he must rest *(22)*.

In the palatability testing of cereal products, from 2 to 5 samples have been served at one time *(66, 159, 203, 204, 217)*. McCammon, Pittman, and Wilhelm rarely served more than 10 egg samples in one day *(187)*, and because of taster's fatigue so large a number as 24 vegetable samples is not recommended *(30)*. In testing dairy products, it was suggested that by increasing the number of samples in a given series it would be possible to test the relative preferences for more possible combinations, and to determine the extent to which the concentration of one ingredient determines the preferred concentration of another *(31)*.

Other investigators recommend the serving of samples in pairs. Fourteen pairs of oil samples were used for selection of panel *(104)*; four or five paired samples of frozen foods were served at one time *(126)*. Precautions that were taken in pairing meat samples gave great advantage when the data were subjected to statistical analysis *(76)*.

Investigations on eggs suggest that duplicate samples be served *(181, 259, 296)*. Gaebe *(122)* used 170 pairs of egg samples.

The error of an experiment may be reduced by replication and randomization *(81)*. Two replicates have been recommended for many products: Eggs *(187, 209)*, fats and oils *(168)*, and apples *(137)*. Griswold employed 2 to 4 replicates for studies on cherries *(136)*; 2 to 3 replicates have been used with poultry *(117, 295)*, and 3 replicates in studies on vegetables *(44, 46, 56, 231)*. The number of replicates for cereal products ranged from 4 to 15 *(130)*; Miller and Beattie used 6 replicates in a study on frozen cake *(204)*. The use of from 2 to 10 replications of dairy products and numerous check experiments are reported in the literature *(10, 62, 178, 280)*. Miller, Lowe, and Stewart *(205)* used 5

replicates in egg studies and Jordan and Sisson *(163)* used 9 replicates. Poultry studies require replication, the number depending on the divergence of the variation produced by the treatments *(184)*.

Use of reference standards

Reference standards have been used by a number of investigators *(64, 65, 85, 89, 124, 152, 182, 188, 193, 201, 227a)*. Tastings are more accurate when made against a definite standard *(85, 148)*; these standards are especially important with persons lacking experience in scoring *(153)*. The use of a standard reference sample of predetermined score may also serve as an unknown to check reliability of judges *(227a)*.

A sample of freshly baked cake was presented with storage samples *(203)*. Handschumaker *(140)* used a scale of five controls for comparison of each sample of soybean oil; others have reported the use of only one control sample *(13, 80, 89)*. A commercial sample was used as a standard for frozen peaches *(165)*.

The use of reference standards in grading dairy products is reported by several workers *(119, 155, 225, 226, 283)*. Downs *(105)* reported the use of three reference standards representing high, low, and medium quality of milk. Jack, Tarassuk, and Scaramella *(160)* used butter made from the normal or regular supply of milk as a control sample.

The use of reference standards is common practice in judging the quality of eggs by taste or smell *(3, 18, 35, 36, 39, 98, 99, 116, 141, 187, 194, 209, 226, 248, 253, 259, 271)*. Sometimes fresh egg is used as a known reference *(3, 39, 184)* and at other times as an unknown reference *(248, 253, 271)*. Known diluted experimental samples are also used *(39)*. In dried egg studies both fresh and dried egg standards are employed *(35, 36, 99, 116)*. Standard reference samples representing specific palatability scores *(259)* are prescored by experienced tasters *(194)*. Some have found it desirable to have fresh shell eggs (48 hours old) included in the tests for purposes of standardizing the judging panel *(4)*.

Many investigators have reported the use of reference standards in judging meat *(61, 157, 233, 260)* and poultry *(90, 107, 184, 185, 244, 263, 264, 295)*. Noble and Hardy *(216)* considered it advisable to judge control samples in a study dealing with possible deterioration in meat quality. Stewart, Hanson, and Lowe *(263)* reported that a fresh control sample of poultry was included in every comparison.

Many have recommended the use of a standard reference sample in judging the palatability of legumes *(30, 97)* and of roots and tubers *(43, 55, 71, 120, 134)* as well as for other vegetables *(44, 276)*. In judging roots and tubers, the method used in grading for each factor against visible standards gave close agreement between scores of different judges *(55)*.

Amount of information given panel

Judges should be informed as to the object of the investigation *(22)*. The panel should hold a period of discussion previous to taste testing. All members should be in agreement as to the weight to be given the various factors to be ascertained and a coordinated approach to their scoring should be established *(63, 275)*.

It is also valuable to have discussion after the test, as it maintains the taster's interest and enthusiasm *(188)*. Discussion after judging tends to favor discovery of important values and to develop better interpretation

of observation among the judges *(89)*. If objectives of the study are not provided, the interest and attention of judges are divided to such a degree that it becomes very difficult to draw any conclusions from the results *(22)*.

Scheduling of samples for concurrent testing

Generally, the narrower the range of samples the greater will be the efficiency of the panel *(69)*; the order of tasting is also important *(148)*. There is evidence that after one has undergone adaptation towards sweetness, one becomes increasingly sensitive to saltiness, making them opposites, as in complementary colors *(202)*. It is common knowledge that a taste is intensified by contrast *(100)*. Sensations excited by different samples should fall in succession and should never be simultaneous *(69)*.

Between tests, the order of cheese samples was changed so that the subjects were uninfluenced by any previous test *(246)*, and poultry samples were randomized so that every treatment could be compared with every other treatment *(263)*.

Beakers containing oils were coded in such a manner that those having the least odor and flavor would be tested first *(213)*. Most panel members preferred to taste oils in their order of increasing flavor *(80)*.

Order of presentation of vegetable samples was changed from time to time but not in a truly random manner. Samples treated with benzene hexachloride often impart an after-taste, so that these samples were always given the highest numbers of the day and judges were asked to rate samples in numerical order *(43)*.

In judging sweet corn, the unknown sample was outstandingly superior and preferred when compared with two low-preference groups, while it was considered among the least desirable when compared with a high-preference group *(101)*.

Correlation of Sensory Tests with Chemical and Physical Tests

Use in interpretation of sensory tests

Cereal products. The protein content of bread showed good correlation with judges' scores *(94)*. There was no correlation between the texture scores of cake and volume measurements *(204)*; however, cakes with greater ability to absorb water were preferred in consumer acceptance tests *(268)*. There was a definite relationship between pH of wafers and cookies and retention of flavor, except in the case of coumarin *(125)*. The changes in flavor of ration biscuits were not as pronounced as changes that were detected by objective measurements of peroxide oxygen values of the extracted fat, fluorescence, and pH, although no attempt was made to correlate the objective and organoleptic data *(196)*. In toast, there was no correlation between percentage of moisture loss during toasting and the total flavor score or any factors of the total score *(67)*.

In macaroni, the correlation coefficients between cooked weight and semolina protein, cooked weight and tenderness score, and semolina protein and tenderness score were all below the 5-percent level of significance *(149)*. Wheat damage, determined by separating kernels according to "light" and "heavy" damage, consistently lowered tenderness score and reduced cooking weight *(149)*.

Beverages. The rate of gas loss of coffee correlated with loss of flavor *(232)*.

Dairy products. Rate of heat penetration, standing index, syneresis, and firmness of baked custard correlated closely with subjective tests *(62)*. Milk of high acidity possessed more oxidized flavor *(10)*. There was general agreement between bacterial counts of milk and palatability scores, depending upon storage times and temperatures *(92)*.

Dried milk fluorescence values showed no significant association with palatability scores. It is unlikely that fluorescence is a good measurement of milk quality *(222)*. Fluorescence of dried milk and egg mix also showed low correlation with palatability scores *(226)*. There was a significant relationship between titratable acidity and palatability, although it was not considered a satisfactory method of predicting eating quality *(224)*. The other objective tests studied (peroxidase value, color intensity, peroxide values, solubility index, colorimetric value, diacetyl value, and fluorescence) were considered unsuitable because many factors were tested simultaneously *(224)*. There was general agreement between peroxide values and palatability, but this may lead to some inconsistencies in comparative results. There is evidence that peroxide formation is a fairly reliable test for relative keeping quality, but this does not apply to gas-packed samples, the keeping quality of which can best be determined by organoleptic tests *(132)*.

In stored frozen sweet cream, there was no direct relationship between pH and flavor *(283)*. A high positive correlation was shown between the initial titratable acidity and the intensity of the oxidized flavor development upon storage when cream was pasteurized at 150° F. for 30 minutes, but there was no correlation when cream was pasteurized at 165° for 15 minutes *(283)*.

In ice cream, a correlation was shown between a low bacterial count and a high total score minus the bacterial count *(198)*. Phosphatase tests showed a tendency to agree with palatability scores, and a relationship was shown between butterfat content and quality scores *(198)*. An increase in milk solids and fat was accompanied by a decrease in size of the ice crystals and a simultaneous increase in smoothness of texture *(72)*.

Butter fluorescence values showed good agreement with palatability scores and were considered valuable in assessing the keeping quality *(129)*. Close correlation also was shown between the degree of acidity of the fat and rancid flavor criticism of butter *(160)*.

Eggs. Ether-soluble fluorescence ratings for dried eggs showed a close inverse relationship to the palatability scores *(181)*. There was reasonable agreement between organoleptic ratings for flavor and fluorescence values *(121, 194, 209, 273, 274)*. In dried egg and milk mix, however, the correlation between fluorescence and palatability scores was so slight that it was discounted *(226)*. There was a direct correlation between the solubility of dried egg powder and the quality of pound cake and Madeira cakes made from the powder *(139, 267)*. The solubility index of plain cake also agreed with palatability scores *(11)*.

In general, there was agreement between palatability scores and results of the following tests on products made with dried eggs: Height measure of custard with penetrometer; volume measure of foundation cake and popovers; pH of yeast rolls; and consistency of mayonnaise *(96)*. There was agreement between the penetrometer readings of baked custard made with dried eggs and the judges' scores on relative firmness *(11, 163)*.

The solubility index of baked custard also showed agreement with palatability scores *(11)*. The results of water-absorbing ability tests, area of slice measurements, and weights of plain cake made with dried eggs showed good agreement with palatability scores *(11)*.

Bacterial count, pH, moisture content, beating value, water value, and KCl value showed a general lack of correlation on prime quality samples of dried eggs. On residue material from secondary dust collectors, moisture content and bacterial count were not associated with palatability; some correlation was shown between palatability and beating value, pH, and water value; there was a higher correlation between palatability and KCl value *(273)*. Results of peroxide oxygen determinations did not coincide with palatability scores for dried egg *(274)*. The correlation was satisfactory between the flavor score of dried whole egg and H_2S tests *(259)*.

There was little correlation between candling and flavor scores of oil-treated eggs in a study of quality changes occurring during storage *(271)*.

In raw egg yolk, there was no significant correlation between differences in flavor and color, although the flavor of dark-colored yolks tended to be somewhat less desirable *(187)*.

With pasteurized eggs, there was some agreement between penetrometer values and palatability scores on stiffness of custard *(141)*. Except for scores on uniformity of texture, there was no agreement between palatability scores and physical tests on cake *(141)*.

Fats and oils. Good correlation was shown between deodorization of oil samples and panel taste scores *(245)*. Peroxide oxygen values, saturated and iso-oleic acid determinations, and iodine number determinations did not correlate highly with the flavor scores *(179)*.

In lard, peroxide oxygen values and alpha-dicarbonyl values gave high associations with odor scores *(128)*. Fluorescence values, although highly associated with odor scores, had regression values too high for prediction of odor scores lower than 5.0 *(128)*.

Desirability of flavor of fats showed little if any relationship to peroxide oxygen values, iodine numbers, and free fatty acid determinations *(168)*.

In shelled pecans, no consistent correlation existed between the chemical test and the flavor ratings for rancidity *(45)*.

Fruits. Apples were judged less desirable and the color became more yellow and less intense as the storage period increased *(137)*.

There was little correlation between the amount of vitamin C in oranges and palatability ratings or acid determinations *(145)*. Palatability ratings by students showed higher correlation with laboratory determinations than did the ratings by adults, suggesting keener taste in the younger group *(145)*. Palatability ratings showed a close negative association with acidity in month to month changes, but not in different rootstocks *(144)*. A definite positive association between palatability ratings and total solid measurements *(143, 144)* and total acid measurements was shown *(143)*.

Judges' scores on cherries indicated an association between an increase in acidity and less desirable color. Objective color data, however, showed that hue changed little with an increase of acid, but became more yellow with large amounts of lemon juice. Chroma decreased as acid was increased, probably explaining low ratings of judges *(136)*. Specific gravity tests showed that desirability of palatability factors increased at first with increasing sugar concentration and then declined *(136)*.

Determination of the darkening index of dried apricots showed that as the darkening increased, acceptability decreased *(258)*. In general, highly sulfured apricots were preferred to those lightly sulfured *(258)*.

Meats. Mechanical shear tests of beef showed high correlation with organoleptic ratings for tenderness *(23, 40, 79, 138, 186, 219)*. In one case, however, flavor scores and quality of juice progressed upward with increased amount of fat in beef roasts, but tenderness scores decreased, which was not consistent with the mechanical tests for tenderness *(42)*. In another case, roasts from calves scored slightly higher in texture than those from yearlings or 2-year-olds, although dynamometer measurements of tenderness showed that they were slightly tougher *(215)*. Histological studies showed significant correlation between size of bundles and texture and tenderness scores *(40)*.

The judges' preference for steer beef was directly related to the ether-extracted content of the edible portion *(42)*. In one study, peroxide values of the fat in beef were indicative of considerable fat oxidation, but much higher values had been expected as taste tests indicated that the fat was almost inedible because of rancidity at this stage *(233)*. High correlation was shown between the mechanical shear test and collagen content and tenderness score *(186)*. There was no apparent trend to distinguish between roasts of high and low calcium content in regard to quality or quantity of juice *(186)*. Moisture losses of 5 percent caused little decrease in palatability scores *(260)*. Palatability scores showed highly significant differences for lots stored under atmospheres of low, normal, and high concentration of oxygen, with uniform decreases as oxygen concentration increased *(260)*. Beef roasts with high phosphorous content had consistently higher palatability grades *(186)*.

Protein fractions showed considerable variation in differences between lambs of various pairs. Nonprotein nitrogen values and intensity of flavor showed no differences *(14)*.

Fluorescence values of dehydrated pork agreed with palatability scores and may prove useful as a measure of quality of this product *(222)*. Highly significant correlation was found between the judges' scores for tenderness and the mechanical shear test *(68, 242, 252)*. Observation of palatability studies showed that the physical state of the tissue of frozen pork changed as the storage period increased *(252)*. Peroxide oxygen values were not associated with any noticeable off-flavor in dried pork *(222)*. The fatty fraction of dehydrated pork may show very high peroxide oxygen values without detection of rancidity by taste panels *(223)*. Peroxide oxygen measurement, when applied to fat extracted from dehydrated pork, was not suitable as an indication of eating quality *(221)*.

Frozen pork roasts stored at higher temperatures had lower palatability scores and higher peroxide values *(127)*. The peroxide development never reached a point that could be regarded as indicating rancidity of the fat, which was confirmed by palatability tests *(252)*. Press fluid determinations showed no correlation with flavor or aroma of the meat *(242)*. The correlation between press fluid determinations and judges' scores on juiciness was highly significant, but the correlation was too low to make it practicable to predict judges' scores on the basis of the press fluid determinations *(147)*. Ether extract of pork showed no correlation with quality of juice *(242)*.

In smoked meats, there was no distinct relation between length of maturation period and flavor quality, although the bacterial count,

peroxide oxygen content of the fat, and color changed consistently with maturation time *(294)*.

Miscellaneous foods. There was no correlation between bacterial counts and palatability scores of frozen precooked meals, primarily because the initial bacterial counts were low in all cases *(126)*.

Poultry. With longer storage of chicken, microscopic tests showed greater disintegration of muscle fiber and connective tissue, and judges' scores on tenderness were higher *(142)*. With aging of fowl, both the mechanical shear test and the taste panel indicated increased tenderness *(185)*. There was no consistent relationship between pressometer values from press fluid determinations and palatability scores for juiciness *(185, 295)*.

Chemical tests for aldehydes and peroxides showed little or no correlation with the evaluation of flavor changes; however, odor alone showed a higher correlation with the chemical tests than the combined score of odor and flavor *(286)*. In one case, peroxide values were paralleled "to some extent" by changes of flavor of the flesh *(107)* and in another case, loss in flavor scores closely paralleled increases in aldehyde and peroxide content of internal fat *(244)*. Iodine number did not vary directly with the flavor scores *(12)*. The effect of storage temperature on palatability was significant with respect to desirability of flavor, but there was no relationship between weight losses during storage and palatability *(150)*. Induction period measurements were not paralleled by palatability scores; nor was there any relation between acidity and any other factor *(244)*.

Vegetables. There was a highly significant correlation between the alcohol-extracted pigment of vegetable greens and either palatability or color as scored by the judges *(231)*. The mechanical shearing test of turnip greens showed that the toughness of the raw greens increased with age although the eating quality was not related to the stage of growth *(291)*. In spinach, the fact that all enzymes may not have been destroyed by the steam blanch process probably accounted for lowered amounts of ascorbic acid and less desirable flavor *(285)*. Ascorbic acid retention in broccoli agreed with palatability scores *(285)*.

Organoleptic tests of beans showed little agreement with qualitative determinations of catalase activity and semiquantitative determinations of peroxidase activity and pigment content *(172)*. There was fairly good agreement between tests for iodine and organoleptic tests *(172)*.

Sugar content of peas showed satisfactory agreement with palatability scores *(30)*. Texture tests of skins and cotyledons and palatability ratings showed only a small number of inconsistencies *(34)*. The relation between taste scores and texture measurements expressed as penetration or crushing values should not be considered final judgments but should be an aid in interpreting variations in mechanical values *(33)*.

In potatoes a definite correlation was shown between palatability tests and carbohydrate analyses *(71, 270, 272, 299)*. Increase in sugar content resulted in a general lowering of quality *(299)*. The proportion of starch to water was correlated significantly with texture score *(272)*. There is a definite relationship between high starch content and mealiness, although this correlation is not perfect and there is no chemical explanation of the noncorrelating cases *(71, 270)*.

Highly significant relationships were found between palatability and color tests *(135, 299)*. Some agreement was shown between specific

gravity tests and taste tests in selection of mealy potatoes *(270, 272)*. There was considerable correlation between the consistency of dehydrated potatoes and the specific gravity of the potatoes from which they were made, but a negative correlation between specific gravity and flavor *(55, 56)*.

Microscopic examinations showed that the size and distribution of starch grains seemed to affect mealy qualities *(272)*. A high correlation between the hydrogen-ion concentration of tuber tissue and the degree of blackening was shown by pH tests *(255)*. The relationship between the oxidation-reduction potential of the tissues and blackening of the cooked potatoes was not consistent *(255)*.

Negative correlation existed between high percent of nitrogen and high quality *(71)*. High percent of dry matter was correlated with high starch content and mealiness *(71)*. There was little, if any, correlation between peroxidase values and quality retention of dehydrated white potatoes *(60)*. There was insignificant correlation of viscosity with mealiness, and no regular relationship between mealiness and gelatinization volume of starch *(270)*.

In mashed potato powder, development of an off-flavor was accompanied by the absorption of oxygen and oxidation of the fat *(48)*.

Desirable taste in carrots was definitely correlated with high refractive indices, when averages of all lots were considered *(46)*. No direct relation was established between extent of carotene loss and off-flavors or between loss of ascorbic acid and loss of palatability *(276)*.

In dehydrated tomato flakes, lower palatability scores were associated with higher moisture content when samples were stored at a higher temperature *(108)*.

It was demonstrated that for cabbage, canned peas, and potatoes, cooking methods can be used which result in maximum retention of both palatability and vitamins *(114)*.

Primary tastes. The pH of the saliva of the judges tested did not appear to have any effect on the tasting ability of the individual *(175)*.

Results of an attempt to correlate sourness with titration against a phosphate buffer checked closely except for tartaric acid *(112)*. However, a buffer titration method is not reliable when other substances, such as salt and sugar, are present *(112)*.

The sweetness index (the ratio of the solubility in water of a sugar to the solubility in water of sucrose) showed excellent agreement with organoleptic ratings for sweetness *(212)*.

Significance of correlation

Flavor estimation is obtained by determining chemical and physical factors for which a high degree of correlation has been established with subjective flavor factors *(6)*.

Organoleptic determination of quality of dried milk powder was a more precise measure than any objective test *(224)*. Hanson, Lowe, and Stewart emphasize the danger of indiscriminate use of objective tests *(141)*. The palatability test as a quality measure of dried eggs for general purposes is the most important test of all. No one chemical test can detect all defects which might be present *(4)*. Organoleptic tests on meat are not a temporary substitute for chemical or physical tests, but must be placed alongside orthodox analyses *(18)*. Organoleptic tests were used in correlation with physical tests to determine the smoothness of

chocolate, which depends on the distribution of the larger particles of sugar in the melted suspension. Results indicated that organoleptic tests could be used to standardize the scale of micro values for milling control and be incorporated into a quality control scheme *(199)*. There are no adequate objective tests that can replace subjective ratings for aroma and flavor of poultry *(184)*. Taste tests are essential in choosing the best methods of processing dehydrated vegetables. They need to be supplemented by chemical tests but cannot be replaced by them *(275, 276)*.

It was considered that the test for alcohol-extracted pigment could be developed as an objective method of measuring quality of vegetable greens because of the high correlation with palatability and color *(231)*. In a test that measured the sour taste of some acids by the amount of phosphate buffer solution needed to bring the pH of the sample to 4.4, it was concluded that in some cases the taste test could be abandoned entirely *(251)*. In another experiment with acid, it was found that acids more sour than 0.0100 M HCl and less sour than 0.0010 M HCl could not be compared by taste measurements *(21)*. Low correlation between press fluid determinations and judges' scores on juiciness of pork showed that judges had an absolute standard of juiciness, or of juiciness and the factors associated with it *(147)*.

Design of Experiments for Food Quality Studies

Choice of statistical design

Cox describes the following experimental designs *(81)*:

(1) Randomized block design — experimental units arranged in groups, each of which contains enough material to form one complete replication.

(2) Latin square design — carries the idea a stage further by grouping the treatments into replications in two different ways to allow for consideration of an additional restriction imposed by the experimental material.

(3) Factorial experiment — method of investigating simultaneously the effects of a number of different factors.

(4) Confounded designs — used when the total number of treatments in a factorial experiment is so large that enough homogeneous material for a complete replication cannot be assembled. Confounding is the arrangement of treatments in blocks that are smaller than a complete replication; that is, each block contains only a fraction of the total number of treatments.

(5) Split-plot design — within the whole plot the subplots receive a second treatment.

(6) Incomplete block design — has for its objective elimination of the heterogeneity of the experimental material to a greater degree than is possible by use of complete block design.

Incomplete block and complete block design have been employed in meat experiments *(219, 294)*, randomized blocks for an experiment on vegetable greens *(133)*.

Importance of proper design

There should be set up a statistical design in order to measure all

variables separately and together and to establish the significance of the results *(102)*. The experiment must be capable of being considered a random sample of the population to which the conclusions are to be applied *(81)*. Error is reduced by choosing an efficient experimental plan *(81)*.

Importance of replication

Replication decreases random errors associated with the average effects of any treatment and will therefore increase the precision of the experiment if precautions are taken to avoid nonrandom errors *(81)*. When tests were made on only one beef animal, definite statements concerning an entire class could not be made *(219)*. Replication should be made, the number depending on the divergence of the variations produced by the treatments *(184)*. Figures reported on a single season's work on corn were considered preliminary and not given statistical significance. Although agreement with other studies on the same variety may be obtained, it is necessary to repeat the test in subsequent years or in a different region to determine the nature and degree of stability of differences between varieties *(49, 101)*.

Simplification of experimental design

When little information is available on the mode of action of any dependent variable, it would be wise to establish these relationships first by experiments of simple design. This may be done by limiting each study to only one dependent variable (such as tenderness of meat) and only one independent variable (such as oven temperature) *(77)*.

Efficient use of time and material

The problem should be analyzed completely before experimental work is begun; procedures should be tested under comparable conditions, then measurements or observations made which describe the effects of each procedure. Error can be reduced by randomization, refinement of technique to secure uniformity in the application of treatments, and the taking of supplementary measurements to help predict the relative performance of the experimental units under treatment *(81)*.

Factors that made tests on dried vegetables difficult were the great variety of vegetables, the difficulty of giving material of graded quality, and the fact that different types often needed to be judged differently *(275)*.

Methods of Analyzing Data

Averages

Arithmetic averages are the most common way of analyzing palatability data. The references are too numerous to list here. Averages have also been used to measure the consistency of each judge *(217)* or inconsistency in scoring successive samples of the same material *(156)*.

Range

A comparison of average ranges is preferred to other methods for

preliminary tests of consistency because it is more rapid and easy to comprehend *(217)*.

Percentages

Results can be expressed as a percentage figure *(65, 202)* as reported by Cameron in a study of the relative sweetness of sugars in which the percentages of right and wrong answers were calculated *(59)* and in grading meat where the number of judgments for tenderness in favor of each roast was expressed as a percentage of the total judgments *(195)*. In an experiment on eggs, each egg had a possible maximum value of 4 points; this figure multiplied by the total number of ballots cast for any one lot of eggs would mean, on the percentage basis, a 100 score for that lot. But if any judge rated one or more eggs less than 4, the final figure for the lot would be less than 100 *(227)*. In the judging of coffee, samples of unknown freshness and staleness are given a percent rating corresponding to the cup in the standard series which it matches most closely in flavor. Then a complete designation of such a sample might be 40 percent fresh and 10 percent stale *(232)*.

Percentage of "direct hits" was not necessarily a good criterion by which to rate a judge's ability in judging milk, since a high percentage of samples scored within a narrow range should yield a larger number of identically scored samples upon rescoring. Rather, the comparison between percentages of all samples scored within the zone and the percentage of all samples rescored with no deviation, would seem to give a better picture of the consistency of the judges' scoring ability *(284)*.

Ratios

Ratios between the total number of answers and number of right answers were used in primary taste tests *(59)*. Variance ratios may be used as an index to measure discrimination and consistency of judges *(217)*.

Chi-square

Statistical treatment of the results on paired judging may be made either by the binomial method or by the chi-square method. The results should be the same whichever method is used. Differences are recorded +1 under treatment preferred and zero under the other. If no difference is observed +0.5 is recorded under each treatment. A chi-square below 3.841 is not significant; above 6.635 is usually considered highly significant for 1 degree of freedom (*77, 255a*).

T-test

Analysis of palatability data by the *t*-test has been reported in the literature *(17, 62, 96, 137, 152, 213, 252)*.

Analysis of variance

This method of analyzing data from palatability tests has been widely used *(11, 16, 17, 40, 97, 104, 119, 134, 135, 141, 144, 145, 147, 155, 156, 163, 166, 168, 172, 180, 183, 196, 213, 216, 219, 220, 223, 224, 225, 226, 229, 231, 242, 244, 252, 260, 261, 263, 264, 273, 274, 294, 295)*. Analysis of variance ascertains the validity of scoring by comparing average scores

(288). Analysis of variance is also used to obtain accurate information about reliability of each judge, his consistency, and his discriminating ability in judging *(217)*. In primary taste tests, analysis of variance was applied to scores of an entire series and showed significant discrimination between groups of samples *(156)*.

Regression

Data from studies on dairy products *(129, 293)*, eggs *(39, 273)*, fats and oils *(128, 168, 213)*, meats *(216)*, and vegetables have been analyzed by regression equations.

In an egg study, the percentage of correct judgments at each of six dilutions used was plotted against the percent dilution. A regression line was plotted for the six data points. A perpendicular line dropped from the point where the regression line intersected 75-percent-correct horizontal line determined the score. This method is adaptable to any food substance which can be made homogeneous *(39)*. Judging characteristics of tasters on primary tastes can be evaluated by computing the correlation coefficients and regression equations relating their assessments to the average of those of all other members of the same panel *(156)*.

For vegetables, regression of flavor on storage time was calculated. Use of regression obscured any variation in rate of deterioration during storage, but results showed no consistent evidence of such variable rate of deterioration *(48)*.

Correlation

Palatability scores have been correlated in many instances with results of other subjective and objective tests *(15, 30, 33, 40, 55, 71, 79, 98, 112, 128, 129, 135, 143, 144, 147, 149, 156, 172, 179, 181, 213, 222, 224, 231, 234, 246, 247, 252, 273, 284, 286, 297, 300)*. For contestants in dairy judging, correlation has also been determined between grades on scores and grades on criticisms *(293)*. However, the ordinary method of testing a panel by the criterion of correlation in trials by duplication was considered questionable *(82)*.

Use has also been made of a modified form of the usual linear type of correlation, adapted for use with the ranks of two variates instead of their actual values. Some extension of the rank-correlation procedure, with respect to its error, was required because of the relatively small values for n *(82)*.

The rank correlation is of service as a quick method of gaging relations between variates which are not normally distributed and when the number of observations is small *(95)*. Any scheme of evaluating placements must take into consideration the relative placement of the individual sample within the series, rather than the sum of their differences in rank *(284)*.

Standard deviation

Standard deviation is a measure of existing variation that is commonly employed in food studies *(17, 33, 34, 47, 60, 61, 65, 75, 76, 113, 119, 127, 152, 156, 180, 181, 194, 202, 221, 227a, 248, 253, 257, 274, 284, 287)*. Standard deviation has also been employed to determine whether the variance of individual judgments of palatability, as expressed in numerical scores, is uniform over the range of quality encountered, and also whether some materials are productive of greater disagreement between individuals than others *(156)*.

Control chart

Application of the control chart method to organoleptic testing is useful for selection of a good panel, for determining what specific taste scores must be examined, for minimizing losses due to failure to pool results of "good" tasters and to pooling results that should not be pooled. It is also useful in the training period, to indicate the length of time training should continue, and in the grading of food quality objectively according to a fixed standard *(193)*.

The control chart was employed for the panel of tasters chosen and for data on experimental samples of eggs *(99)*, dairy products *(225, 283)*, fats and oils *(213)*, and primary tastes *(84)*. The control chart method can help in the selection of those tasters whose judgments are valid as defined by the control chart for averages, and reliable as defined by the control chart for standard deviations *(194)*.

Over-all ratings

To arrive at a sound over-all rating or score for a sample from ratings of a number of properties or components is often quite difficult, especially if unrelated or independently variable properties are to be considered in the over-all rating. Simple arithmetic means, however weighted, are often quite unsatisfactory. One characteristic may render the food totally inedible and yet the total score may be relatively high. One solution is to arbitrarily make the total score zero whenever the score for any one of several critical components is zero. Another solution is to set a minimum acceptable level not only for the total score but also for each of various critical components *(65)*.

By using the geometric or harmonic mean, the over-all quality so derived is more representative and more in accordance with direct judgment than the arithmetic mean *(162)*. A maximum of, say, 10 points is awarded for each of n characteristics and the product of individual scores divided by $10n-1$ to reduce the result to a number of reasonable magnitude. Thus, any one characteristic which renders the foodstuff completely inedible (score zero) results in a total score of zero also *(162)*.

Plank *(228)* suggested a method for determining total food quality which included subjective evaluation, objective physical or chemical measurements, biological value, and external state. In calculating the total grade, the same "weight" cannot be awarded to all properties. Thus, he introduced "specific multiples" for each factor and multiplied the grade number with the corresponding "specific multiple." The arithmetical evaluation of the total subjective grade is much facilitated if the sum of all values of the different "specific multiples" equals 10. The grade "zero" was used only to express the unconditional rejection of a sample.

Discriminant functions

The use of discriminant functions enables one to compare a composite of several variables pertaining to one "method" or "treatment" with a similar composite pertaining to another. A discriminant function composed of scores or decisions on texture, flavor, aroma, moisture, and appearance gives one value based upon the five variables or measurements; this single value, made up of five values, can then be compared statistically with another single value based on similar measurements pertaining to

the second recipe. Instead of having to test each factor separately for significance, a discriminant function provides a means for testing a combination of all measurements *(16)*.

Missing values

In an investigation by Hardy and Noble *(147)* on pork loin roast, in every series one judge was absent at least once; therefore, in each series the ratio of his scores to the average scores of the other judges was determined for each time for which he was present, and the scores for those times when he was absent were calculated from this ratio.

Application to food products

Beverages. Beer: The significance of the results of the triangular test can be calculated according to Bengtsson's adaptation of the chi-square analysis *(151)*.

Cereal products. Method of ranking in order of preferences was applied to cake *(47)*.

Dairy products. Cheese may be ranked in order of firmness *(246)*. Transformation of simple ranks into numerical values by use of Fisher and Yates' table 20 was used by Bliss, Anderson, and Marland *(31)*. In transforming simple ranks to scores such as 1, 2, 3, and 4, distribution departs more from normal form than is desirable for analysis of variance. First and last choices tend to be ranked more easily than intermediate items in the series. Fisher and Yates' table 20 corrects this tendency.

Eggs. For the calculation of a flavor index, or a number that will agree closely with a flavor score given by a competent panel to a spray-dried whole egg (or any type except one contaminated by foreign matter), an equation based on several chemical determinations is offered by Fryd and Hanson *(121)*.

Fats and oils. Grant and Lips *(128)* suggested prediction equations and errors of estimate for assessing rancidity in lard. Lemon, Lips, and White estimated storage life of oil statistically from data of the panel. In a few cases graphical interpretation was necessary *(179)*. Panel ratings of soybean oil were normalized and converted into scores which were summed for analysis *(140)*.

Meat. Ramsbottom and others *(233)* tabulated the number of judges who indicated preferences and those who showed no preferences in a study of frozen beef.

Poultry. Harshaw and colleagues *(150)* tabulated descriptive terms as to the number of favorable and unfavorable comments on flavor. Wills *(295)* used weighted scores combining judges' scores and objective tests.

Vegetables. Greenwood and Salerno *(133)* changed ranks to scores, using Fisher and Yates' scores for original data. Stillman, Watts, and Morgan *(266)* assigned numerical weights to arrive at a composite palatability score.

Dove *(101)* reported that preference placements can be analyzed according to their distribution. One test showed four types of distribution: (1)

Skew to right or to left, representing high-preference varieties and low-preference varieties; (2) bimodal in part, representing varieties ranking high in one test and low in another; (3) equal distribution, showing completely unorganized behavior; (4) normal or near normal distribution.

Caldwell, Lombard, and Culpepper *(56)* averaged all grades for potatoes, then expressed them as numerical ratings on a scale in which the sample found to be best in any particular factor was rated 100 for that factor, and others were rated from this standard. Also, each sample was given a final grade on general desirability, which was a summation of all scores on all factors considered. In determining the final grade, the factors of flavor and texture were each given twice the weight for color. Samples were also designated as excellent if average numerical grade was 90–100; very good, 80–89; good, 70–79; fair, 60–69; poor if below 60.

Primary tastes. Results of primary taste tests were expressed as the geometric means of the frequency distribution of the molar solution of the respective substances tested *(112)*. In another case, quinine alkaloid was assigned a value of 100 and the relative bitterness of other products was calculated from it *(243)*.

Significance and validity of results

Analysis of results is not easy. Different panel members will exhibit different preferences. Results should be analyzed statistically in order to secure a satisfactory assessment of relative values of products being tested *(209)* and to determine the reproducibility of judges, individual panels *(184)*, and groups of panels *(4)*. Summaries based on numerical interpretation of descriptive gradations of flavor are valid only to the extent that proper weights have been given to the terms describing the flavor *(253)*. There should be analysis of scores obtained from preliminary panel testing and if necessary, factors should be changed before further scoring is made *(63, 64)*.

The principle of treating scores as if they were true numbers is wrong. They must not be treated as numbers unless they are actually shown to possess the properties of arithmetical numbers, such as $2+2=4$. The danger of allotting numerical values to what are frequently nothing more than categories, is not sufficiently realized *(18)*. One should guard against the error of attributing significance to small differences between palatability grades *(9)*.

No advantage is gained by adding together the marks given separately for color, flavor, and texture in order to get an over-all assessment of quality. Totaling marks obscures the judgment on specific characteristics. Failure to obtain an adequate standard in respect to any one criterion is sufficient to render a product unacceptable *(276)*.

Results need to be summarized, displayed, and interpreted. The theory of probability is used to find out how much confidence to place in the results and what sense can be made of the figures. An efficient, flexible set of statistical tools is available, but an understanding of the basic assumption involved in their use is necessary *(81)*.

Although a difference does not reach the arbitrary level of significance generally used in statistical analysis (5-percent level), this does not imply that no difference exists; it merely indicates that the difference, be it real or not, is smaller than the experimental error *(213)*. The main sources of error in an experiment are failure to standardize the experimental technique and the inherent variability in the experimental material *(81)*. Selection of the experimental error is important *(81)*.

PROCEEDINGS OF CONFERENCE

ESTHER L. BATCHELDER, *Chairman*

Methods of Measuring Differences in Food Quality

Discussion[3]

ELSIE H. DAWSON: The methods of measuring food quality reported in the literature can be divided into about six basic methods—with many modifications in actual practice. The difference is sometimes slight, yet sufficient to make comparison of results from different laboratories impossible. The most widely used methods are scoring, ranking, paired comparison, triangle or triple comparison, dilution test, and the use of descriptive terms.

Descriptive terms are usually accompanied by numerical scores, although they are sometimes used alone and the results analyzed as percentage of judges who noted certain characteristics. This method is often used in preliminary work to find out what characteristics are important. However, the results of such tests are difficult to analyze and report.

Numerical scoring is perhaps the most popular method and the most difficult and misused method. Scales ranging from 1 to 5, 7, or 10 are most commonly used, although they may range from 3 points to 100 points. Whether the scale should be restricted to 10 or 5 grades or extended to 50 or 100 depends on the taste sensitivity of the judges. A 10-point scale has been used by many investigators but the interpretation of the scale may vary greatly; sometimes 10 is perfect and in other cases 1 is perfect quality.

The advantages of using a 5-point scale with only the highest and lowest points defined are given as (a) it avoids much of the difficulty of devising adequate description of flavors and (b) it is difficult to achieve linearity when every point is defined.

At the Taste Testing Conference at the University of North Carolina (Nov. 7–10, 1949), J. W. Hopkins of the National Research Laboratories in Ottawa, Canada, reported on a universal scale applicable to many food products and any characteristic:

+5 gross excess	−1 very slight deficiency
+4 very decided excess	−2 moderate deficiency
+3 decided excess	−3 decided deficiency
+2 moderate excess	−4 very decided deficiency
+1 slight excess	−5 gross deficiency
0 ideal	

One criticism of the scoring method is that too many variables and too many characteristics are usually included. An individual must have considerable training before he can obtain consistent results in grading.

In the ranking test, judges are asked to rank samples in decreasing or increasing order of some characteristic. Ranks may be converted to scores, using table 20 in Fisher and Yates' book, Statistical Tables (1949). The ranking method encourages judges to make fewer distinc-

[3] The chairman designated one person to open the discussion on each subject considered during the conference. She also called on those who had previously indicated that they had material to present on certain topics, but otherwise the discussion was informal. See page 132 for identification of persons participating in the conference.

tions among samples and reduces the tendency of the judge to prefer certain score ranges.

In the paired test, two samples are submitted to the judge and the judge is asked, "Which is more tender?" "Which is sweeter?" and so on. Sometimes a standard sample is presented first and the judges are asked which of the two unknowns is the same as the standard.

In the triple comparison (triangle) test, three samples are examined, two of which are duplicates. Judges are asked if there is any difference among the samples, and if so, to select identical samples. Both the paired and triple comparison tests are useful in selecting a panel but costly for use in an elaborate experiment.

The dilution test determines the smallest amount of unknown that can be detected when mixed with a standard material.

Other methods of testing quality of foods include matching with standards and weighing or measuring amount of food left on plates.

I have not gone into details as speakers and discussion to follow will elaborate on sensory methods of measuring differences in quality.

DAVID R. PERYAM: I want to mention the bibliography (Taste Panels) which we prepared in 1947 with about 400 titles on palatability testing. It is available without charge to anyone who requests it. I think we have about 100 copies.

I am going to present some methods which we use for evaluating differences in our laboratory, and also a test showing the good results we are getting with these methods. We call them "discrimination" tests, because they utilize the ability of persons to discriminate among foods. One of the tests—the triangle test—has already been referred to, so I won't spend much time on it. Also, I will not discuss the statistical computations used to analyze the significance of difference, since that belongs properly under the statistical section of the conference.

In regard to scales, we have used J. W. Hopkins' 11-point scale, which has 5 points going in one direction and 5 in the other, and functions as a double 6-point scale (Biometrics 6: 1–16, Mar. 1950). We have also used a 9-point scale, with 4 points in each direction and have found it to give better reproducibility than the 11-point scale on preference evaluation tests. We have found this to be true with large groups of people in the laboratory. We have no evidence at this time that the 9-point scale is the best for the judging situation where you have trained individuals judging quality, but we are working on that problem. J. P. Guilford (Psychometric Methods, 1936) mentions a study involving 23,000 judgments, in which the 9-point scale resulted in the highest reproducibility. This happened to be a study in which people judged the ability of other persons. They also make the point that 9 points are about all that people can handle anyway.

Of the difference methods which we use, one—which is similar to the paired difference—we call the duo-trio test. We use three samples with one labeled as a control. Then we have an unknown pair, one of which will be a control, with the order determined by chance. The person doing the testing is given the control sample first, and then at controlled time intervals, the other two samples. He is asked to select the sample which is different. The significance of difference is arrived at in terms of the number of correct answers. In this situation, half of the answers may be right by pure chance. In using this test on milk, we find that we can give two or three tests at one time, but we know we can't give as many as five.

Preliminary to this testing, we use a "warm-up" to get the flavor in

the subject's mouth. All the samples, after the first one, are tasted in a background of flavor from the preceding sample. There's nothing you can do to change that situation so we give the "warm-up" sample with the instructions to taste it and forget it in order to make the situation the same for all samples. We use time intervals of 10 or 15 seconds, but have some evidence that with some foods like milk, 5 seconds or less would be better. We felt that we shouldn't go below 10, but some people feel that they can do better when they taste the samples rapidly. There is, of course, the danger of getting a blend of flavors if the samples are tasted at too close intervals, or you may get confusion from prior stimulation. There is also the factor of forgetting; some people can remember for 30 seconds while other people can't remember for 5 seconds.

In addition to the duo-trio test which I have just described, we use also the triangle test, which is fairly standard. There is not the control with this method that there is with the duo-trio. The three samples are offered at the same time, with no controls on the amount tasted at one time or the time intervals between tastes. However, the method serves just about as well as the duo-trio and is much more convenient for the tester. Of course if you are serving hot foods, the temperatures are not the same when each sample is tasted. Even with these disadvantages, we find that this method gives good results and is even superior to the duo-trio for many foods, despite the lack of controls on temperature, quantities, and time intervals. We are doing further research on these methods in our laboratory.

The third discrimination test is a paired difference for odor testing which is somewhat similar to the duo-trio. We take advantage of the fact that recovery from odor stimulation is quite rapid, and present two standards. These are offered for a controlled length of time, usually two or three sniffs back and forth. The person is instructed to smell them until he can detect a difference and has established a criterion, and then he is given two unlabeled samples for identification. We call this the dual-standard odor test. We have used it for taste, too, but find that for taste the single-standard method is better.

SYLVIA COVER: We developed the paired-eating method to test differences in tenderness for our meat cookery work. Roasts are not homogenous and so we obtained our paired samples by matching a small portion from the right side of the carcass with a similarly placed portion from the left side. These two samples were presented to a judge who was asked to record which sample was the more tender. This method may be used for testing odor, flavor, or other characteristics. We found no significant difference in tenderness of roasts cooked by the same method. But when they were roasted by different methods (oven temperatures of 225° versus 125° C.) the results showed a highly significant difference.

Further tests with beef, pork, and lamb showed a range in tenderness percentage from 51 to 96. The higher tenderness percentages were associated with the slower rates of heat penetration, but no method had at that time given meat which was always very tender. Then we further decreased the rate of heat penetration by using an oven temperature as low as 90° C. in a drying oven versus 125°. Some of the roasts cooked at 90° took 48 hours to cook well-done. The increase in tenderness was again highly significant and the roasts were at last judged very tender.

CLAUDE H. HILLS: I am going to discuss the use of a simple scoreboard for testing, originally used by Washington Platt. There is one pictured in E. C. Crocker's book, Flavor (1945). It is a sheet of white cardboard

with a scoring scale of 10 to 1 on the left-hand side. At the top, there are numbers which represent the different samples. If you are scoring a food on a rating scale, you place the standard sample on the appropriate number. For example, if the standard rates 7, you place it on 7. The judge tastes the standard first, then the samples to be rated, placing each on the appropriate place on the scale, according to his judgment. The judge doesn't have to have a pencil in hand when he scores. Also, he can conveniently retaste the samples that rate close together and either place them in their original place or change them.

After the preliminary tasting, the judge may wish to take a cracker or a drink of water, and rest before retasting the samples which were close together in his original evaluation. He may decide now that 9 and 8 should be reversed, because the order in which samples are tasted does affect the judgment. We believe that the scoreboard cuts down on fatigue since it reduces the number of times you have to taste a given sample. It took me a couple of weeks to get our people to use the scoreboard, but after using it once, they continued to use it.

HELEN MOSER: What do you do about recording the scores from the scoreboard? Also, what happens if your people do not get to compare their results?

CLAUDE H. HILLS: We have mimeographed sheets and the samples are coded. After placing their samples, the judges record their scores on the sheets. I think it is good technique not to let the people discuss their results.

DAVID R. PERYAM: Have you tested the use of the scoreboard for reproducibility? In other words, have you gotten around to all those important things most of us don't get done?

CLAUDE H. HILLS: There are people here better qualified than I to go into the merits of rating scales, pairing tests, etc. I have presented the scoreboard simply as a convenient technique which enables the judge to concentrate on the job at hand and not to have to write while he is judging the samples.

GERTRUDE COX: Have you thought of this as a ranking method to which you are adding a score refinement?

CLAUDE H. HILLS: It is a sort of combination of ranking method and scoring method, in which you score the samples, then go back and try to rank the close scores. I think you will agree that when you have samples that score closely together, one of which you tasted at the beginning of the test and one later, you should go back and compare them again because in the interval you have tasted other samples which affect your judgment. We analyze the results as scored data, which I believe gives you more information than if analyzed as ranks.

RUTH JORDAN: After having tried out some other methods of evaluating differences in flavor, we decided to use the dilution method for our experimental work in tracing the effect of storage on the flavor of dehydrated eggs. We were interested primarily, not in whether one product was preferable to something else, but in whether or not change had taken place during a period of storage. We needed a method which would require a relatively small number of tasters and one which could be used at different periods of time.

The method decided upon was to put a given percentage of the dehy-

drated egg, which was our experimental sample, into a sample of egg that was strictly fresh. The standards came from the same lot of hens and were at all times under 24 hours old. Believing that the sensory perceptions are based on a definite degree of intensity of stimulus, we established the standard for our dilutions on a percentage of decrease basis. That is, we started with a score of 1 for 100 percent dilution, with the next step 20 percent below that, and the next 20 percent of that, each step being 20 percent below the preceding one. In the end, we had dilutions ranging from 100 percent for a score of 1 to 1.4 percent as our end point for a score of 20. We had set up the brackets in which these dilutions might be detected. We ultimately set up a system whereby we had 12 trays or 6 dilutions with duplicates for any one period. We did preliminary work with our panel to be sure that after the judges had gone through the 12 trays, they could still perceive differences to the same degree of precision.

We always had before the person one sample which was marked fresh and the other sample which was marked experimental. On that same tray, there were eight unknowns, which could be an equal number of experimental or an equal number of fresh samples, or they might be seven and one or other combinations. The judge, after comparing the experimental and the fresh controls, listed which of the coded samples were experimental and which fresh.

If the sample was such that the difference could be readily detected, we got a high percentage of right answers. If the percentage dilution was such that difference could not be readily detected, we got a number of guesses or a number of wrong answers.

In compiling the results, we used six points in plotting a curve, placing the percentage of right answers on one axis and the score based on the percentage of dilutions on the other axis. Since we needed to find some point at which we could say a decision had been reached, we decided on that point at which the number of 75 percent right answers crossed the line. From this point, we dropped a line down to the axis on which we had plotted the scores.

By this method, we were able to determine over a considerable period of time that there were changes that took place in some of our samples. All of the data were treated for dependability and our judges were tested periodically for their reliability throughout the testing period. This method would be adaptable, of course, only for those products which can be made homogeneous. In our tests, the materials were mixed thoroughly and the eggs scrambled so that there was a homogeneous mass.

J. C. HENING: I am going to say a few words about score cards. The consensus at the Raleigh meeting was that it is a good practice to use some type of score card and that the value of the score card depends somewhat on the experience of the tasters in using it. In Geneva, we use a score card with a vertical numerical scale of 1 to 10, and across the top we designate the sample numbers. We use 10 as denoting excellent; 9 as very good; 8, good; 7, good minus; 6, fair; 5, fair minus; 4, slightly poor; 3, poor; 2, very poor; and 1, extremely poor. We don't have all those descriptions on the score card, but we have them posted in the room in which we taste. We have all become accustomed to this type of score card and use it for all of our testing, even the triangular test where two samples are alike.

I am going to describe an experiment we carried out on canned peas for the purpose of determining the desirable stage of maturity of the

peas. There was a comparison also of the effect of different fertilizers on the peas. We used the Thomas Laxton variety from five different harvests grown on four plots of ground. At each harvest, we had four different samples of peas to taste. We used a scale of 1 to 10 with the sample numbers across the page.

The peas were served at room temperature just as they came from the cans. We tried heating the peas, but had no evidence that we got any better results that way, so we served them without heating since it was more convenient. We used 3 cans of peas from each harvest for the purpose of obtaining a uniform sample. The peas from the first harvest were very young and we got a divided opinion from the group. Some of them rated them high and others scored them down.

Tenderometer readings were made when the peas were harvested, and the panel scored the flavor, texture, and color of the canned peas. There was high correlation between tenderometer readings and the subjective scores as shown in Exhibit 1.

Where different fertilizer treatments were used, we paired the fertilizer treatments for each harvest. The panel varied in number from 10 to 15 at different times and was made up of both experienced and inexperienced tasters. Some of the scores fell quite a way out of line and it may be that they should have been eliminated, but we included all of the scores in our results. In this test of the comparative maturity value of peas, the average deviation of the scores of the judges was less than one point from the mean scores.

EXHIBIT 1. ORGANOLEPTIC RATINGS AND TENDEROMETER READINGS
THOMAS LAXTON PEAS

6–12–6 fertilizer
1949

Sample No.	Tender-ometer reading	Flavor	Texture	Color	Harvest
1339	91.5	6.4	7.8	7.0	First harvest, 4 lots.
1340	93.0	6.6	7.9	7.0	
1341	93.3	6.7	7.8	7.0	
1338	93.8	6.0	7.7	7.0	
1345	99.25	7.1	7.5	7.0	Second harvest, 4 lots.
1344	100.25	7.3	7.2	7.0	
1343	102.00	7.1	7.0	7.0	
1346	104.5	7.0	7.2	7.0	
1342	109.25	6.9	6.8	7.0	Third harvest, 4 lots.
1349	112.0	5.8	5.7	7.0	
1347	115.0	6.7	6.5	7.0	
1348	116.8	5.8	5.5	7.0	
1350	129.5	5.6	5.4	7.0	Fourth harvest, 4 lots.
1353	134.3	5.6	5.2	7.0	
1351	138.0	5.7	5.1	7.0	
1352	142.8	4.2	4.0	7.0	
1355	165.5	4.3	3.9	7.0	Fifth harvest, 4 lots.
1356	167.0	3.9	3.8	7.0	
1357	168.0	4.3	4.4	7.0	
1354	172.8	4.0	3.8	7.0	

BERNADINE H. MEYER: I will take only a few minutes to say something about our experience with score cards for evaluating quality of precooked foods in freezer storage. We have worked with a 1 to 5 scale, 5 representing excellent or very good down to 1 for very poor, in evaluating the quality of each characteristic. When we want the over-all or cumulative effect of several qualities, we have arbitrarily prorated the values in multiples of 5, assigning 5, 10, 15, 20, or 30 points for any one quality, making the total score 100 points. The judges continue to use the five gradations of quality in their scoring. Two sample score cards are shown in exhibits 2 and 3.

EXHIBIT 2. JUDGING RECORD FOR FRUIT PIES

Date_____ Name of judge_____

Directions: Place the number corresponding to the term which best describes the food in the proper column. Use multiples of 0.5. Write any comments you may have in the column and space to which they refer.

Recipe					
Storage time: Sample No.					
Acceptability 5 Very good (in all respects; you know of no improvement) 4 Good (enjoyed it; minor improvement desirable) 3 Fair (could eat it without enthusiasm; improvement needed) 2 Poor (edible, but that is all) 1 Very poor (inedible)					
Color of fruit 5 — 4 — 3 — 2 — 1 Very good to Very poor					
Texture of fruit 5 — 4 — 3 — 2 — 1					
Flavor of fruit 5 — 4 — 3 — 2 — 1 Characteristic to Off or absent					
Quality of pastry 5 — 4 — 3 — 2 — 1 Tender not soggy to Tough and soggy					
Flavor of pastry 5 — 4 — 3 — 2 — 1 Very desirable to Very poor					

Remarks:

EXHIBIT 3. JUDGING RECORD FOR SPONGE CAKES

Date_____ Name_____

Directions: The possible score for each cake is 100 points. Place score in columns to right. Use whole numbers. Write any comments at bottom of page or in column with score. Give full value for excellent quality; 4/5 for good; 3/5 for fair; 2/5 for poor; 1/5 for very poor.

Sample: Storage time: Sample No.	1	2	3	4	5	6	7	8
I. General appearance — external_____								
1. Shape_____10 Regular, slightly rounded and free from cracks.								
2. Volume_____15								
3. Crust_____ 5 a) Tender — not too smooth, sticky or crusty. b) Color — light brown, free from spots and without a moist shiny appearance.								
II. Appearance — internal_____								
1. Grain_____10 a) Cells — small, uniform and thin walls. b) Free from large air spaces. c) No compact layer.								
2. Texture_____15 a) Tender, moist, feathery, light in weight to size, not compact or soggy.								
3. Color of crumb_____ 5 Light yellow — not gray or off color.								
III. Flavor_____								
1. Taste_____30								
2. Odor_____10								
Total score								

Comments:

Cake 1.

Cake 2.

Cake 3.

Cake 4.

Cake 5.

Cake 6.

Cake 7.

Cake 8.

Perhaps I can point out more of the limitations than the advantages of these score cards as we have used them. One of the shortcomings lies in the fact that we need more information from the judge than just the score. If the grain of a cake is scored down, for example, we need to

know whether the reason is because of irregularity in grain, too compact grain, or because the cell walls are too thick. Scoring without any explanation does not give us enough information. There isn't always enough room for the judges to write on the score cards.

QUESTION: How did you determine the weights of the various factors for the total score?

BERNADINE H. MEYER: We made arbitrary decisions as to the comparative importance of the various characteristics in the over-all quality.

HELEN J. PURINTON: At our institution we have no center for palatability testing, and since five departments are interested in such tests for different reasons, we found it necessary to develop several types of score cards. With a brand new product we prefer the 1 to 10 scale described by Mr. Hening. We use it too in studying variations in methods of preparation. We do considerable work on squash in our part of the country because it is grown so extensively, and whether it is baked, boiled, or prepared in some other manner, the 1 to 10 score card seems to fit our needs.

We use this score card also for testing and training new judges. We give them a standard consisting of a sample considered highly acceptable, and get their reactions to it as a test of their sensitivity to taste and smell. We then use the score card for evaluating the product.

For testing fruits and vegetables we usually make use of a very general score card with provisions for rating appearance, aroma, and palatability; the latter characteristic is subdivided into texture and flavor. We have an over-all rating point system of 1 to 100 and we weight it according to our interest at the time. For example, with a new variety of squash that is in the process of being developed, the chief interest would be in its appeal — the eye appeal, especially color — and we would weight appearance heavily. If there is no question of acceptability on the basis of appearance, and a problem has developed in the cooking process, we might weight heavily some such factor as aroma or palatability which we break down into texture and flavor. We frequently use 35 for appearance, 15 for aroma, and 50 for palatability. These scores vary according to what the grower wants to know. A sample score card is shown in exhibit 4.

EXHIBIT 4

Product_____
Date_____
Name_____

1.	Appearance		*35*
	Color	15	
	Eye appeal	20	
2.	Aroma		*15*
3.	Palatability		*50*
	Texture	20	
	Flavor	30	

I consider this product
( ) excellent
( ) good
( ) average
( ) poor
( ) unpalatable

We have developed another score card with 6 points for poultry meat judging. A sample score card is shown in exhibit 5. Our section of

EXHIBIT 5. SCORE CARD FOR POULTRY MEAT

Name_____ If excellent — Score 1
If average — Score 2
Date_____ If poor — Score 3

	Score	Remarks	Score	Remarks	Score	Remarks	Score	Remarks
1. Color								
2. Aroma								
3. Flavor								
4. Tenderness								
5. Texture								
6. Juiciness								

the country is engaged in developing the broad-breasted chicken which we evaluate on the basis of six factors: Color, aroma, flavor, tenderness, texture, and juiciness. We use a scale of 1 to 3 since we are evaluating general acceptability only in this preliminary work. A separate score card is used for white and dark meat, with an interval between judging times for each. We use what might be called a quadrangle test, in that we judge four birds at each panel. Three of the birds are alike, of a standard variety on the market; the other is the experimental sample.

The same group 2 or 3 days later tests another four birds, of which three are experimental and one of a standard variety. Our panel does not exceed 12 in number. The results of the two judging sessions of the same panel are then correlated. These methods have worked out to the satisfaction of our Poultry Department, which is primarily interested in the testing.

In testing potatoes, our Agronomy Department uses a byproduct for taste testing. Actually, they are interested in potato varieties that are adaptable to our long winters and short summers. Practically everyone likes potato chips so they have no difficulty in getting a panel to judge in midmorning or midafternoon by nibbling a few potato chips. The potatoes are routinely made by the same process into potato chips. We have used practically the same panel for the past 5 or 6 years.

We have experienced some difficulties in testing the desirability or adaptability of different varieties of strawberries for preservation by freezing. We have decided that we want a judgment on just the flavor, so we have to minimize other differences such as size and color. There is also the factor of differences in the rate at which different varieties thaw upon removal from the freezer. If one variety of berry is icy and hard when tasted, and another soft, it introduces another factor into the judging of flavor. So when we are after only flavor in judging the berries, they are blended in the Waring blendor and the judges taste the puree. This method has another advantage in that the samples can be sweetened equally by addition of a sirup made with a given weight of sugar or sweetened with a given weight of sucrose.

MARY L. GREENWOOD: We use a very simple score card in detecting off-flavor in potatoes, which perhaps isn't really quality rating at all. We have one column for our samples and another for rating the intensity of the off-flavor. If there is none, the judges indicate it with a minus sign. If it is there in some given degree, they indicate that with a plus mark; if in some greater degree, with a double plus, and so on.

MILDRED BOGGS: At the Western Regional Laboratory, we have three different groups of people working on taste testing, if we wish to call it that. There are about 6 persons in each group so there are 18 of us doing full-time taste testing.

On our score cards, we try never to use the words desirable, attractive, excellent, poor, but we try rather to describe the characteristics of the factor to be scored. We do not weight scores for the various characteristics into one score for the sample. The purpose of all our work is to determine the causes of deterioration and how much deterioration has occurred under some method of processing, storage, or other treatment as compared with the strictly fresh product. We are not therefore interested in any over-all evaluation, but rather in knowing such things as how much the color or flavor has changed.

L. C. CARTWRIGHT: First, I want to say that some of you have seen this paper, Organoleptic Panel Testing as a Research Tool. I will pass around the 40 copies I brought with me. This paper attempts to promote organoleptic panel testing in a much broader sense than for just testing food palatability. I am suggesting its use as a research tool in the laboratory to supplement chemical and physical methods of evaluation of any properties that affect sensory response, particularly the senses of odor, taste, and smell. This includes appearance also.

We use descriptive terms primarily in consumer tests, tests with

untrained panel members, and in training work with panel members. If we set up a numerical scoring system for evaluating particular properties of a food product, we discuss with the panel members the meaning of the numerical scores in descriptive terms and try to arrive at a consensus of all panel members as to just what the numerical score means in terms of a given degree of quality. For trained panel members, we find a saving in time and money without any sacrifice in accuracy in the use of a simple score sheet with numerical scores only.

In testing a product which the manufacturer wants to get into the hands of the consumer in the same condition in which he produced it, we use the freshly manufactured or processed product as a standard. We score the samples that have been aged or stored in various types of containers, such as paperboard, envelope, or carton, for varying lengths of time, against that standard sample. In accelerated aging, to evaluate the effect of certain containers on a product, we control carefully the temperature, humidity, and other conditions and use the same conditions for aging the standard food product as for the experimental sample. In this type of test, we use both an aged and an unaged control. We divide the score of the experimental sample, which has been aged in contact with the package material in question, by the score of the aged control to get what we call the material score for the packaging material.

In regard to ranking and paired tests, we use ranking with descriptive terms sometimes on special problems where we are examining only a few samples. Paired tests are used similarly if there are just a few samples to be evaluated. However, we use both the ranking and the paired tests regularly in connection with our ordinary testing, using numerical scores.

We train the panel member to go through the samples, usually five or six, and rank them by arranging them in sequence and assigning numerical scores. This is on the scoreboard plan, but without using a scoreboard. Then, if there are some close samples, he goes back and uses the paired comparison. He really uses the blindfold method, in that he turns the coded samples so he cannot see the markings and shuffles the samples. He then retests these two samples, and if on two or three trials, he ranks them in the same order, there is no question about his evaluation.

We score odor, flavor, and aftertaste. Aftertaste is very important on certain food products. The odor scoring is done first, followed by flavor and aftertaste. The interval between tasting individual samples varies from 30 seconds to as much as 2 or 3 minutes, depending on the nature of the food and the individual taster. It takes usually about 15 to 45 or 60 seconds to get the full value of the aftertaste and to be able to evaluate it, and then a little period to get the normal flow of the saliva, depending on the food product. We may use water between samples, and we may not. The most important thing is to allow the normal flow of saliva to return. Such allowance should be made after using a water mouth rinse because there won't be a normal taste response until the saliva flow is normal again.

The three components, odor, flavor, and aftertaste, are broken down into two components each for scoring. This is done for two reasons, one purely statistical and the other functional. These two components are presence of desirable and absence of undesirable characteristics, which are scored separately. Statistically, we find this gives us greater reproducibility, which we believe is due to the fact that we get more accurate evaluation because panel members think in terms of desirable and undesirable characteristics of the food product. We have applied

this method of evaluation to a wide variety of products: Powdered prepared cocoa mix, fruit beverages, coffee, tea, cocoa, alcoholic beverages, spices and spiced foods, chocolate candy, and potato chips, to name a few.

The most typical application of this method is to packaged foods which have picked up flavors from the packaging materials. There it is more a matter of evaluating the intensity of the undesirable odor and flavor. We get highly consistent results with very slight individual panel member deviation when things are running normally. However, when we run into an unusual off-odor or off-flavor, as for instance ink odor in packaging material, or varnished or lacquered paper with a little of the solvent odor left in it, or some other characteristic that the panel hasn't been encountering, we get wider scattering of results, requiring rechecking. We may then use the triangle or paired test. Dilution methods have been used only for flavor and odor intensity. Dilution sometimes alters the quality and must be used with extreme care.

MILDRED BOGGS: We also score intensity of two components of flavor. We call them characteristic flavor and off-flavor. The system works all right except with samples exhibiting considerable off-flavor. Then we get large variation in scores for characteristic flavor. No doubt this results from the fact that we have not decided how to train judges on this point. We simply do not know whether a sample can have pronounced characteristic flavor when there is quite a lot of off-flavor present.

L. C. CARTWRIGHT: It depends on the food product and on the nature of the off-flavor. We would like to be able to separate them entirely. In practice you can't, but we find our panel is fairly consistent in separating them. Using coffee as an illustration, suppose you take good fresh coffee and add even 1 percent of a very stale and rancid coffee to it. I doubt if anyone would be able to judge that coffee as full-bodied in its desirable characteristic. I think the presence of the undesirable characteristic would make it impossible for you to recognize the full-bodied flavor. It is a compromise, without question, but we believe, based on our experience, that it is better to use that method of judging than not to use it.

COMMENT: We use that method and have lots of trouble with it. In general, we believe that if you have much off-flavor, you may as well not score desirability and that it is only where off-flavor is slight that it makes any difference whether you score both or not.

L. C. CARTWRIGHT: That may very well be.

QUESTION: Do you find that these two tend to add up to constant score? That is, presence of desirability and absence of desirability.

L. C. CARTWRIGHT: No, I don't think they do. Our method of scoring is precisely equivalent to scoring components separately on a fixed scale, say 0 to 10, and then taking an arithmetic mean for your total score. In the case of odor, flavor, and aftertaste, they are so closely interrelated that we believe it is sound to take an over-all score. We find that it saves time and is more convenient to do our weighting beforehand and merely add up to 100 for maximum score, which is in effect taking an arithmetic mean. Our standard weighting there is 15–15 for odor, desirable and undesirable, 20–20 for flavor, 15–15 for aftertaste, giving a maximum score of 100. On that basis, with a panel of eight experienced tasters on prepared cocoa mix, we found that over a long period of time,

we generally got a standard error of the mean of not greater than plus or minus 2, which on scores that range between 75 and 95 is very good. We get average deviation of individual panel members from 3 to 6, with only occasional deviation outside that range.

We have an arbitrary rule of throwing out all scores of a panel member who deviates from the panel average more than 20 on the basis of 100, or who misses the coded standard control by more than 5 points; in other words, scores below 95. Since we have six components, he can't even score it off 1 unit out of 15 or 20 on all six of the components and have his scores accepted. Incidently, in the last 6 months, I don't think we have thrown out anybody's score on missing the standard control, and only about two or three (which is less than 1/100 of 1 percent or so) on missing the panel average by more than 20 points.

COMMENT: We ran into the problem in frozen asparagus that the texture rating appeared to be influenced by the off-flavor. That's something hard to get around.

L. C. CARTWRIGHT: Yes, those are factors you have to contend with. We have run into a lot of these outside variables. I keep an eye out for those things as the data flow across my desk and if I see any unusual results, we go back and check up to find the cause. Very often we find that it is some factor like that, which has cropped up unexpectedly, and it may require reexamination of that series.

COMMENT: I would like to comment on your scoring for desirable and undesirable characteristics. I think perhaps it helps solve a problem we have had in regard to how much weight to give an undesirable characteristic. Sometimes a panel member believes a product to be very desirable but finds a little bit of undesirable characteristic there, and marks it very low, on the theory that there shouldn't be any undesirable characteristics at all. On a scale of 10, he might mark it 3. Another panel member might consider the slight undesirable factor relatively unimportant and rate the product 8.

L. C. CARTWRIGHT: Previously, we ran into that same problem, and it led us to adopt our present rating system. I might mention also that there is a space on our score sheet after each sample for remarks. Ordinarily, our trained panel members do not make comments. If they have scored a sample down, particularly if they have scored it down for some unusual characteristic, they will write a comment, which helps us in analyzing the results. These comments explain unusual deviations in scores.

CLAUDE H. HILLS: What is your basis for choosing weightings of 15 points, 20 points, etc?

L. C. CARTWRIGHT: The weights were chosen arbitrarily, based on general experience.

CLAUDE H. HILLS: We have had a problem in determining quantitatively the concentration of a volatile fruit essence. This, of course, has nothing to do with detecting an off-flavor. We also have some storage tests where we want to detect loss of quality. We have other methods, too, of checking loss of quality. The test I shall describe is an attempt to compare two or more samples of a fruit essence for strength. What we have done is to run a threshold dilution test with a series of samples, each decreasing in strength so that it is 50 percent of the previous sample. This gives a twofold qualitative difference between successive samples.

We arrange the beakers containing the samples in rows. The panel members first smell a beaker of distilled water. They then smell the weakest sample, then the distilled water again, and then the next sample. When they arrive at the one in which they can just barely detect the fruit odor, that is considered the threshold dilution for that person. We used this method with 11 people and obtained a statistical distribution of the sensitiveness of these people. You may be interested to know that with six samples we found the confidence limit to be about 30 percent. In other words, the limit of detection or accuracy of the method is plus or minus 30 percent. This is perhaps a unique illustration of a dilution test.

HELEN MOSER: This score sheet for oil testing was developed from the score sheet used when we were testing rations for the Quartermaster Corps, U. S. Army (exhibit 6).

Committee Report

Purpose of tests

The purpose of all the methods to be described is to determine differences between samples which have had different treatments. The maximum information which can be obtained includes kind, amount, and direction of differences. This type of test is not related to consumer preference or food acceptance.

Methods of testing

Paired and triple comparison tests. In the paired test two samples are submitted to judges. Sometimes a standard sample is presented first and judges are asked which of the two unknowns is the same as the standard. In the triple comparison, or triangle test, three samples are examined, two of which are duplicates. Judges are asked whether there is any difference among samples and if so to select identical samples. In these tests, direct comparison of the samples requires only short memory. They are relatively easy to do and the statistical calculations are not difficult. Their limitation is in the small number of samples which may be compared at one time. Paired and triangle tests do not give amount of difference except under special conditions. They are not suitable for use in storage tests unless fresh or frozen controls are available.

Dilution tests. By the dilution technique the smallest amount of unknown that can be detected when mixed with a standard material is determined. Results are influenced by taste or smell thresholds of judges employed. The method applies only to homogeneous materials and requires suitable standard material. A necessary precaution of the test is to dilute with the same kind of food material, not with water, that is, dilute milk with milk, eggs with eggs, etc.

Ranking tests. In ranking tests samples are ranked in decreasing or increasing order of intensity of one quality. Ranking may be preferred to scoring in balanced incomplete block design when extreme variation is present.

Scoring tests. Scales in use range from 0 to 10, 1 to 10, 1 to 5, 1 to 3, +5 to 0 to −5, +4 to 0 to −4. Total scores of 100 made up of the sums of several factors with the effective scoring range often considerably less are

EXHIBIT 6

NM–291

Name_____Date_____

Please indicate the score by placing a check mark (√) in the space opposite the proper intensity.

		SAMPLE 1		SAMPLE 2		SAMPLE 3		SAMPLE 4	
		O	F	O	F	O	F	O	F
	10								
Good	9								
Less desirable but acceptable	8								
	7								
	6								
Objectionable	5								
	4								
Unpleasant	3								
	2								
Repulsive	1								

Please indicate intensities of flavors by placing check marks opposite the proper flavor: (√) weak; (√√) moderate; (√√√) strong.

	SAMPLE 1	SAMPLE 2	SAMPLE 3	SAMPLE 4
Bland				
Buttery				
Beany				
Rancid				
Painty				

also employed. Scale length is related to the number of intervals that a judge can distinguish or differentiate. An individual must have considerable training before he can obtain consistent results in scoring and a panel works best with the scale length for which it has been trained. Research is needed to determine the best length of scale and whether the optimum length of scale is affected by the material or the factor under test.

Descriptive terms

Many laboratories conduct taste panels on the basis of preference, using descriptive terms such as excellent, good, fair, and poor to describe quality. The ratings may be established by reference to standards of top, medium, or low quality, or by a preconceived notion of what these might be, that is, an imaginary standard. Several testing groups have adopted the use of terms to properly describe a flavor characteristic, such as "oxidized," "feedy," to describe off-flavors in milk, and "nutty," "grassy," "painty," "fishy," to describe oils. Some laboratories have gone still further and used fully descriptive terms as a basis for panel operation. The panel member studies the test material and becomes acquainted with the problem. Then the panel members together prepare the set of descriptive terms which are to be used and the score to which each applies. In the "flavor profile" method, the character notes for both aroma and flavor are expressed in common terms. During testing, the order of appearance of each character and its intensity are recorded. The amplitudes of total aroma and flavor are also recorded.

Recommendations

For preliminary work many factors such as odor, flavor, juiciness, and tenderness may be used. When the experiment begins, the number of factors tested should be reduced to as few as possible, that is, two or three characteristics on which special emphasis is needed.

Further research is needed (1) to evaluate the various methods of measuring food quality on a basis of purpose, precision, and efficiency of design, time, and material; (2) to develop reference standards for all characteristics of the various commodities and the stability limitations of these standards.

COMMITTEE: Mildred Boggs, Elsie H. Dawson, David R. Peryam, Loren B. Sjöström, Sylvia Cover, Chairman.

Panel Selection

Discussion

MILDRED BOGGS: Exhibit 7 shows the influence of panel size on the standard error of the daily mean flavor score for 94 scrambled dried egg samples. When only five tasters were used they were part of the larger panel and the best tasters of the group. The best tasters were selected on the basis of the correlation coefficient for the first score and the duplicate score. The exhibit shows a considerable decrease in error with increase in panel size even though poorer tasters were included in the larger panel.

Exhibit 8 shows the influence of panel size on significance of differences in several tests with frozen peas. I selected these particular tests from recent data because they showed very small differences between samples and therefore would be likely to accentuate the effect of panel size. The samples were originally scored by a maximum of 15 judges. Absences often reduced this number somewhat. Four or five replications were done in each test and daily means were used in the analysis of variance.

Tests 1 to 6 included three samples, only one pair of which was ever significant. Test 7 contained four samples, three pairs of which were

EXHIBIT 7. INFLUENCE OF PANEL SIZE ON STANDARD ERROR OF MEANS OF DAILY SCORES

(Smallest panel is made up of best judges of the panel)

Standard error	Percent of samples with given standard error			
	5 tasters	9 tasters	12 tasters	14–16 tasters
0.10–0.14	1	5	12	7
0.15– .19	11	24	33	50
0.20– .24	15	37	39	43
0.25– .29	16	21	16	
0.30– .34	23	11		
0.35– .39	14	2		
0.40– .44	7			
0.45– .49	8			
0.50– .54	3			
0.55– .59	1			
0.60– .64	1			

EXHIBIT 8. INFLUENCE OF PANEL SIZE ON SIGNIFICANCE OF DIFFERENCES IN SEVERAL TEST SETS OF FROZEN PEAS

(4–5 replications of each test)

Test No.	Significance level			
	4 judges	6 judges	8 judges	All judges (usually 11–12)
1			*	*
2		*	*	*
3	*	*	**	**
4	**	**	**	**
5	**	**	**	**
6				*
7a	*	*	**	**
7b			*	*
7c				*

* indicates significance at 5-percent level; **at 1-percent level.

significant with all judges. When only four of the judges were used they were the best ones of the group and in case of an absence the fifth best judge was used. The panel of six was the six best judges, and so on. The best judges were selected on the basis of all triangle data available for each judge over a period of 1½ years and usually included 150 to 200 triangle tests.

The exhibit shows that in some instances we picked up slightly smaller differences or increased the significance level by increasing the panel size, but we did not improve our situation as much as might be expected. We would be satisfied with smaller panels than we use if we knew in advance who the good tasters are and that we could count on regular attendance, but since we do not know these things we usually carry about 12 judges on each panel.

Exhibit 9 shows the effectiveness of a 6-week training period. We had one pea panel that had been in operation for about 3 years but we needed an additional panel for this product. We trained a new group for about 6 weeks, selected the best of the group as a panel, then compared this new panel with the old one. The samples for the comparison contained 0, 33, and 67 percent of a poor-quality sample, the remainder being an excellent-quality sample served as puree. These samples were scored four times and each individual's scores were used in analysis of variance, sources of variation being treatment, replicates, and remainder. The mean differences between samples and the pairs which were significant at the 1-percent level, the mean square of the remainder, and the F-ratio are shown in exhibit 9. The exhibit shows that the newly trained panel gave about the same performance as the experienced panel. Apparently

EXHIBIT 9. COMPARISON OF PERFORMANCE OF A NEWLY TRAINED PANEL WITH A PANEL THAT HAD BEEN SCORING ALMOST DAILY FOR 3 YEARS

Judge	Old panel[a]				Judge	New panel after 6 weeks of training			
	Mean difference					Mean difference			
	0–33	33–67	MS_r	F		0–33	33–67	MS_r	F
CH	1.5	0.6	0.8	6	EB	1.7	0.3	0.4	8
RM	1.0	1.8	.5	15	CB	1.0	0.0	.7	2
FL	*3.0	1.0	.3	39	JD	.5	*1.5	.2	19
RR	− .2	1.8	.6	6	LD	*2.0	*2.0	0.0	∞
VS	2.2	1.1	.8	14	BL	1.2	.5	.2	13
WS	2.2	1.0	1.1	10	MP	*1.5	.5	.3	13
AW	1.0	1.8	.5	15	CS	.5	*1.5	.3	13
RW	*1.6	*2.1	.1	157	WW	1.8	1.0	.5	15
EW	1.5	1.2	1.9	4	EW	−1.0	*4.7	.1	163
AW	*1.5	*1.5	.3	27	JW	*1.0	*2.0	0.0	∞
Panel[b]	*1.5	*1.4	.2	47	Panel[b]	*.9	*1.1	.1	37

[a] Minus difference indicates wrong sample received higher score.

[b] Panel results include scores for 4 or 5 judges who were absent too much for analysis of individual results.

* Difference is larger than L. D. at 1-percent level.

training pays, at least in this instance for which we had much experience and knew exactly what to train for.

As to personnel available for panels, we are lucky. There are about 375 people in our building and we can use anyone who qualifies and is willing to serve. We are willing to take as prospects anyone in the building, male or female, any age, smokers or nonsmokers, and so on. We train, select, and then check performance of individuals on routine tests. If they can do the job, that is the only matter of importance to us. We are not satisfied, however, with some of our methods of checking performance, but I will discuss that when we come to it.

We consider interest extremely important in order to get the best results. We find that our tasters like to be right, they like to be consistent and reproducible, so they will take advantage of every solitary bit of information they can garner. We therefore do not give them much information in advance, but keep up their interest by giving them the full results of every experiment after it is finished, as well as their own individual performance in the test.

RUTH B. BOYDEN: I come from a small station where persons available for judging are limited in number. We use staff members usually from the Department of Home Economics, and find that they are interested in the testing. Those teaching food preparation have perhaps a little more interest as well as experience in judging food quality than the teachers of dietetics. The results with these people seem more reliable than those with graduate students in foods. We have not been able to consider age but have felt that health is important.

GLADYS L. GILPIN: I am going to describe the method we used in selecting a panel for judging canned chicken. We had quite a large amount of canned chicken to judge and we had no panel ready. One of the chief things we considered important was for the judges to be able to identify rancidity. Rancidity, which is frequently encountered in canned chicken, is a fair indication of quality of the pack and of how well the pack is keeping. With the help of our statistical people, we devised a little test to determine responses of different people to varying proportions of rancid meat mixed with natural-flavored meat. We used a paired test and ground all of the meat samples. One sample was all natural-flavored chicken and the other was a mixture of different proportions of rancid and natural-flavored meat. We used an adaptation of Abraham Wald's method of sequential analysis for carrying out this test (Sequential Analysis, 1947).

There were six series of samples, with varying amounts of rancid meat thoroughly mixed with the natural-flavored meat. The natural-flavored sample was used as a control throughout the test. We started with rather large proportions of rancid meat and found that the tasters detected it immediately. The first series had 1 part of rancid to 7 parts of natural-flavored chicken. We decreased the amounts in each succeeding series. The last series contained only 1 part rancid to 256 parts of natural-flavored meat. That is a very small amount of rancidity and it takes a person with a keen sense of taste to be able to identify the sample.

If a person had no errors in a minimum of 10 pairs of samples, she was considered acceptable as a judge at that particular level of concentration. If she made a few errors, 1 or 2 but not as many as 4, we would test her with more samples so that the percentage of errors made would put her in either the acceptable or the rejected group. For instance, if she made

1 error in 10, we could not accept her but we still couldn't reject her as she hadn't made 4 errors, so we had to test her further. If she made only 1 error, she could have up to 23 samples to qualify for acceptance. If she made no further errors, she would be acceptable at that level. We sometimes had to offer as many as 30 samples, but everyone tested could be classified either as accepted or rejected, and the sampling could be done fairly rapidly. One of the advantages of this method is the fact that it is economical of time and samples as the test can be terminated for each person as soon as she either qualifies or is rejected.

We discovered that if a person was rejected at a particular level, she was rejected at every higher level. This confirmed the assumption that the point of her first rejection was her threshold for recognition of rancid flavor in these samples of ground canned chicken.

DAVID R. PERYAM: We have set up what we consider to be a practical system for panel selection. The value of the panel member depends not only on basic sensitivity but also on such things as adaptation time, recovery time, the important factor of memory for odors and tastes, also adjustment to the test situation. If you tested all these factors independently, you would do nothing for a half year but select a panel. We have used the duo-trio or the triangle test to compare people's abilities to detect differences in products which are going to be tested. We have set up panels for dried milk, for coffee, for detecting the presence of pepper, and for detecting the presence of monosodium glutamate.

We start with groups of as many as 90 persons with the problem of selecting about 20 for judging. We use the triangle test and set up pairs of samples with arbitrary differences. The differences must not be so large that everybody detects them. If they are detected only as frequently as chance will allow, we must conclude that the correct identifications may be due to chance alone and therefore the test is not accomplishing its purpose. We start with comparatively large differences and decrease the differences as we progress with the testing. We rank our 90 persons, by giving them a series of 16 to 20 tests each, in order of sensitivity or, rather, discriminating ability.

QUESTION: Do you run four samples or four triangles at one sitting?

DAVID R. PERYAM: No. They usually do two sets at one sitting. Sometimes they do two in the morning and two in the afternoon, if we can schedule it that way. It takes a lot of scheduling for 90 people.

QUESTION: Do you use triangles on the same two samples?

DAVID R. PERYAM: Yes. For example, individual A will eventually take four triangles and he will have a score on the first test anywhere from 0 to 4; then he will take four on test 2 and four on test 3. Sometimes we drop a test. Let's say in test 4 that one person got two out of four samples, another person got one, another three and another none. This is a chance pattern, and we would probably throw the test out and try another.

QUESTION: Does each successive test have less difference between samples A and B?

DAVID R. PERYAM: Not necessarily. For example, with monosodium glutamate tests, we want panel members who can identify its presence in a wide variety of foods. We won't test the effects in any specific food. Maybe the first test will be with beef stew, followed by a test on mashed potatoes, then one on corned beef hash.

To test the reproducibility of the panel member's judgments, we run about 17 to 20 replications and calculate the coefficient of correlation for each individual. We can then select those individuals with high correlations. We haven't determined how high a correlation we are going to demand. Actually in the one problem where we used this check following the sensitivity tests, we were very unhappy to find that some of the people selected on the basis of sensitivity could not reproduce their judgments very well. We decided to leave them out, but to use them when we were testing only for difference.

QUESTION: I'm speaking for those of us who don't have 90 people to draw from. It would appear that if you didn't have people at hand who were both sensitive and able to reproduce their judgments, you would have no business being in palatability testing. Am I right?

DAVID R. PERYAM: No, I wouldn't say that. You would devise a system for evaluating the people you have. Even with a small number, you can evaluate the individuals. For example, if you have someone who establishes a correlation in replicate tests of below 40 or maybe below 50 on samples to be tested, you know that he is not much good. Actually, these discrimination tests don't tell you how good a person should be, but they do compare the persons tested. You might find that out of 10 available people, 3 are so far below the others that you will have a better panel if you drop them. Selecting from 90 people does offer the opportunity to select the cream of the crop. It also means a whole lot of work in testing.

HELEN J. PURINTON: We are in one of those places where there is difficulty in getting panel judges. There are available at the most around 10 or 12 people. We have panels no smaller than 10 and we try to include 3 or 4 people who have done routine food testing for at least 1 year. We have worked out a training plan for our judges and will use only those who score 60 or more on the training tests. In retesting an old product, we will keep only those individuals on the panel who score 80 or more.

We have just finished testing the 19 people who are available right now. The group includes both males and females and is made up of research people, townspeople, housewives, general students, and student dietitians. Of the 19 tested so far, 14 scored over 75 percent three times, which is considered excellent. They must score 75 percent twice before they can serve on a panel. It would appear that, even with the small number to choose from, we should have some pretty good judges, granting of course that the test itself is good.

We always check personal likes and dislikes of the panel members and never ask a person to judge a food he does not like.

MILDRED BOGGS: We find that a test on judges for one product doesn't do you a bit of good on another product. We have people who completely flunk out on, say, sulfur dioxide and are excellent tasters on some other kind of flavor.

L. C. CARTWRIGHT: I would like to say that we have found some individuals who are generally sensitive and skillful in detecting the differences we set up. They pass the tests for all of the panels.

HELEN J. PURINTON: We feel that it is necessary to test persons for each product.

L. C. CARTWRIGHT: We do that, too, but we do find that some people

seem good on all tests. We are not confident enough of them, however, to accept them without testing for each product. While some people are generally good, even our best panel members without exception will fall down on one or another type of product or evaluation. No doubt, the person with extensive previous panel experience has a higher chance of doing well on a new product. Psychological factors are important in the ability to evaluate odor, flavor, and palatability. The approach, the interest, the training in evaluating, in thinking, in concentrating on the factor to be evaluated — all these are factors in successful judging.

We have found that calling panel members from their usual jobs may result in mental block. Panel members who are usually good may be immersed in a piece of work which is interrupted by the judging and may give judgments out of line on that occasion. They may be careless because they want to get back to the job. We try to fit panel members into the sessions most convenient for them.

We have not made any studies on age or sex of panel members but have not found any relevance there at all within the group with which we have worked. We have considered health important, insofar as it affects sensory perception. We have had panel members ask to be excused from the panel or go completely off the beam on a particular product because of a dislike for that food or a feeling that there was something wrong with the food that would affect their health.

Because of the cost of more elaborate testing, we usually select panel members with one to three replications of the triangle test. Then we train the panel. If we are going to have more than one or two sessions on a particular product, we give the panel a brief training and use a larger panel for the first several sessions. The results are analyzed and those who are in least agreement with the panel average are eliminated. If we and/or the panel members have had no previous experience with the product or properties to be tested, we usually start with 15 to 25 members and feel our way — set up standards and criteria, train, and select our panel all at the same time. On a continuing problem, we may come down to as few as 4 to 8, depending on the precision required or the degree of difference between the samples that we are evaluating.

We worked with as few as four or five panel members in much of our work on spices. Highly accurate work is required on the cocoa product I mentioned. Our client is accepting or rejecting production batches of packaging materials on the basis of our tests. We use eight members on the panel for this work and they have almost never failed to reproduce a score within a 2- or 3-point variation, that is, 2 to 3 percent on 100, with scores ranging mostly between 75 and 85. In general, our reproducibility has been consistent with the calculated standard error of the mean.

QUESTION: How often do you use replications of tests?

L. C. CARTWRIGHT: We often use replications. Some of the work, however, is with either concentrations or products on which we have done previous tests. For instance, if we have tested a certain group of spices in one food product and are testing them in another food product, we may not do a replicate test if our results are consistent with those we have got before. If our results do not agree (and it may be that in a different food product they shouldn't agree) we will run replicate tests. If we continue to get the same deviation from previous results, we will attribute it to the difference in the food product.

GLADYS E. VAIL: We belong to the groups that have a limited selection

of panel members. Most of our work is done on a cooperative basis with other departments on the campus, so if we undertake new projects, we do palatability tests only if the other department furnishes part of the panel. This plan helps us to get a panel and also brings about a better acceptance of the results, since they understand what we are talking about when we turn in our report to them. Without such experience, there is sometimes a tendency to reject the findings. They understand the results better when they have participated in the experiment.

EDWARD TOEPFER: Should the question be brought up as to whether one good judge is better than a panel?

L. C. CARTWRIGHT: I think, generally speaking, that no one judge, no matter how good, is as good as a panel, even a relatively poor panel. My reason for so thinking is that a judge may be highly accurate but every now and then even the best judge will go off completely. I think we should not depend on any one expert, no matter how good.

B. L. RIBACK: I have just one word to add to what Mr. Cartwright has said. Speaking again from the commercial standpoint, when you are making a product for a national market, with wide geographical differences in tastes, it is courting disaster to depend upon any one individual's judgment of the product. Our experience leads us to believe that the best type of panel is one made up of both sexes and having adequate distribution of racial groups, nationality groups, and even age groups. In addition, there must be geographical distribution.

I would like to make the point also, that in the final analysis, we need to judge what the consumer will accept. Perhaps the consumer will not accept as high a quality of product as you want to put out. We need to establish first the consumer's standard of value or acceptance and then judge a product in accordance with that standard.

Committee Report

Factors to consider in selecting a panel include availability and qualifications of the panel members and size of the panel.

Qualifications of members

Panel members must be available for testing. They must have both time and interest in the problem. For these reasons, members of the staff and cooperating groups frequently serve as panel members. Experience is considered desirable, although it does not assure their qualification. Some skills are believed transferable.

Judges' qualifications are usually determined by testing their relative sensory acuity and reliability with the food to be tested. They are often tested also for the primary tastes. Types of tests used include paired tests, triangular, duo-trio, multiple, and dilution. Evaluations are made by (1) ranking, (2) stating preference, (3) identification or differentiation, (4) description, (5) numerical score, and (6) combinations of these methods. Both acuity and reliability vary in different persons, with different products and different qualities, and both vary also from time to time.

There are various opinions in regard to the optimum age for judges. Many consider age largely irrelevant. There is similar variation in opinions regarding sex of judges. Health is considered important. There

is general agreement that judges should be free from colds, infections, allergies, fatigue, and worry. Psychological factors play a part in judges' performance. Good judging requires that panel members have an interest and some understanding of the problem. They should like the food being tested.

Further investigation is needed in many areas. Studies could throw light on the value of experience versus sensitivity; the significance of sex, of health, and of the many psychological factors involved in judging.

Size of panel

A small, well-selected, and well-trained panel is considered more precise than a large, unselected, and untrained panel.

Further investigations are needed to determine whether many replications with a small panel are better than a few replications with a large one. The question of "how small can an expert panel be" should be studied.

It was the opinion of the committee that the size of panel needed to provide given precision cannot be determined in advance. In practice, it is usually necessary to start with a larger panel for preliminary work, analyze the data, and then reach a determination regarding the panel size. If preliminary work is not possible, data may be collected from a large panel in the expectation that some may not be used if analysis indicates that certain judges are not qualified.

COMMITTEE: Ruth Boyden, Gladys L. Gilpin, Maude P. Hood, L. C. Cartwright, Chairman.

Training of Panel Members

Discussion

L. C. CARTWRIGHT: We consider the training of judges highly important. We find that panel members can make important contributions to the training session on how to evaluate the product, set up the score sheet, etc. We discuss the approach to the problem, type and method of evaluation, and come to an understanding in regard to the meaning of the numerical scores and the descriptive terms for each. We have the panel members smell and taste a standard control sample of the product to be tested. We have another sample which has been scored, to illustrate how the scoring is to be done.

As we proceed with the testing, if we find a panel member whose scores are deviating from the group, he is called in for discussion and perhaps reoriented with a number of samples of known score.

One of the checks we use from time to time on our panel as a whole is to give replicates with some interval between, to see if they give similar scores. The training and evaluation of the judges is a continuous process.

MARY E. KIRKPATRICK: We also like to have the judges help us establish standards. I am going to describe the training which we carried on some time ago. The fact that we are working with biological material that does not stay the same influences our methods and procedures. We have to work quickly with the training program for judging potatoes or else the potatoes have changed before we are ready to start testing.

In evaluating quality, we are interested in three types of perception: Visual, which takes in color and appearance; texture, which covers

mealiness and wetness or dryness; and flavor. We asked our judges if they could evaluate the characteristics in that order. Most of them said they could. We consider it important to have a plan of testing which meets the approval of the judges.

Our study was set up for a 6-week period with a judging session each day. We used a 3-point scale on the score sheet which the judges helped us set up. For example, color was indicated on a 3-point scale with a descriptive term for each point. Moisture content, texture, and flavor were set up in a similar manner. If we found that some descriptive term was not understood or agreed upon by the group, it was freely discussed the first day or two and a term which was better understood and agreed upon was substituted. We found that this plan helped to maintain the interest of the panel members.

In our test, we tried to prepare a cooked product that would rate a 3 for color, one which would rate 2, and another a 1 score. That method was repeated all the way through for the four or five characteristics — five characteristics for boiled potatoes, four characteristics for mashed potatoes, and five for baked potatoes. We considered one characteristic at a time. The judges said they liked to study one characteristic at a time.

We studied each characteristic for a week and started the first day with samples intended to illustrate each of the scores, 1, 2, and 3. The samples were discussed freely and no scoring was done. On the second day, we gave the judges three knowns which were samples of the 3, 2, and 1 scores, and in addition two unknown samples. They scored these samples. The plan included offering similar samples on the third day. Sometimes, however, our samples did not fall into the score classes intended. In spite of the controls we used — selecting from known specific gravities and known varieties and locations — we ran into some problems in maintaining uniformity of samples. Our tests were done in the summer and we believe that early fall would be preferable.

Potatoes are not a homogeneous material and one tuber is not a representative sample. It takes six or eight tubers to make up a good representative sample. We used wax models as reference samples for color and for showing sloughing of the outer coat of the potato.

We found that our judges improved with experience. Later tests with the panel showed that the training and judging experience helped these panel members reproduce their judgments.

QUESTION: Did you test for only one characteristic each day or all on one day, but only one at each sitting?

MARY E. KIRKPATRICK: We spent one week on one characteristic for boiled potatoes, the next week on another characteristic for boiled potatoes, and the next week on another characteristic. It is time-consuming and may not be the best way, but it is the plan we used in our study.

MILDRED BOGGS: In general, we train new panels in about the same way as Mr. Cartwright does. Briefly, this procedure includes training with samples that exhibit all the characteristics of interest in the study, agreeing on terminology to describe characteristics, trying to standardize level of scores if this is needed — but I should mention that it is not needed if you analyze differences — then practice and check effectiveness of the training program.

In addition to training new panels, we often need to add new members

to panels already in operation. Here we use regular sessions for training periods. The prospect attends the regular session but is told what the samples are and what to look for. Following scoring we discuss his scores in relation to those of regular panel members. We carry a prospect along like this for a while, then let him serve on the panel just as a regular member but do not use his scores until they measure up to regular panel performance.

HELEN J. PURINTON: We have set up tests based partly on data obtained from Dr. Langwell of the Drexel Institute. We use envelopes containing filter paper impregnated with the different basic tastes. One sheet is plain, one impregnated with sucrose, another with citric acid, one with salt, and the fifth with quinine sulfate. The judges note the degree of taste as excessive, moderate, or slight. For slight, we use the thresholds recommended by Mr. Crocker. We increase this concentration fourfold for moderate, tenfold for excessive.

The judges are put through the tests three times. A judge scoring less than 60 percent is considered unacceptable. We want our judges to score 75 percent or over. We consider scores of 90 to 100 as excellent, 80 to 90 as good, 70 to 80 fair, and 60 to 70 acceptable but poor.

After each test, we expose the judges to four foods which have been previously judged by a panel and see how close they score to the given values already obtained for those foods. We have used only frozen foods so far, but expect to use others. We are also developing odor tests.

GLADYS L. GILPIN: In training judges for our panel on canned chicken we encountered the problem that with processed foods, the very fact that the product is canned means to most people that it has an overcooked flavor. We were interested in evaluating four factors — the amount of natural flavor, the off-flavor, juiciness, and tenderness. Our score card had 5 points. We first tried to find samples which illustrated the different levels on our scale. This was a time-consuming job in itself.

We prepared the samples of canned chicken by different methods: (1) Simmering in water for 1¾ hours, (2) simmering in water for ¾ hour, and (3) heating in broth 5 minutes. We changed the water once during the cooking time for the first two methods. The latter method of heating the chicken for a 5-minute period was adopted as the standard preparation procedure for the experimental samples as the flavor differences showed up best with that method.

We were interested in having the judges recognize overcooked and overprocessed flavors, so we had them score samples that were processed the minimum safe length of time and others that were processed an additional 50 minutes, which produced severe overprocessing.

To prepare samples for tests on juiciness, we heated some canned chicken samples in the oven to dry them out. The white meat responded to this treatment more satisfactorily than the dark meat. We heated some samples in the oven 45 minutes, some for 1¾ hours, and compared them with pieces heated in their own juice for just 5 minutes.

Preparing samples to illustrate differences in tenderness presented a problem. Even with old roosters, the processing tenderized them so much that it was difficult to show differences. We finally used plain cooked samples of old and young chickens to illustrate tenderness differences.

We used the score cards, and through free discussion, agreed on levels which would score 5, 4, 3, and so on. The reference sample was rated 4

on some characteristics, and 5 on others by the group. We continued to have it available to each person during every judging period, using it as a known sample.

We found it took considerable time to prepare the chickens, can the meat, and then judge the samples. The training period was, as a consequence, somewhat shorter and less adequate than we had hoped it might be when we planned the work.

HELEN MOSER: In 1944 our laboratory was confronted with the problem of evaluating soybean oil. It had been customary to run peroxide values on samples to determine whether or not they were edible. We concluded that these tests were not entirely satisfactory and decided that the organoleptic method might solve our problem.

Our laboratory followed the procedure set up by the Bureau of Human Nutrition and Home Economics in giving preliminary acuity tests to about 40 people and selecting a panel of 12, based on the test results. Those 12 people had been screened by two previous tests. This panel operated from about 1944 until 1946 when, for a variety of reasons, only about 3 people were left, and it was dropped entirely. Shortly after it was reorganized, I came into the picture and faced some new and different problems.

After reviewing the literature, I felt that the acuity tests were not related to all tasting and that the correlation was not high enough to warrant continuance of the previous method of panel selection. We asked for volunteers and got 11 people and asked the previous group of 11 to return and run through a test period with us. We set up a training program based on actual oil samples. Although we were going to test soybean oil, we did some control work with cottonseed oil. We felt that they should know mineral oil, too, because many of our dilution tests were made with it.

We used a round-table type of training period. There was a large score sheet on the wall and the trainees were given samples of the various oils. We discussed the samples openly but when it came to flavor, they were asked to describe the flavor on their score sheet. We discovered that some were timid about expressing their judgment in the beginning. We worked with them for a 3-week period in this informal fashion, discussing odors and flavors and giving them an opportunity to score samples. We used the wall score sheet to rate samples so all could see and come to an agreement regarding flavors present and the scores to assign to them.

In order to check their performance, we set up a series of samples which could be evaluated statistically. We presented the new training panel and the panel with previous experience the same samples for scoring, and drew up correlation regression charts for each member. We also ran duplicate samples so we could draw control charts both on the panel average and the individual performance. Then in order to test their ability to distinguish differences, we gave the triangular test. From the results of these three tests, we felt that we had enough information to use as a basis for selecting our panel. Of the 22 persons tested, 18 were selected. Control charts, which have been run continuously through the 2 years that these people have been serving as panel members, indicate that the training was worth while.

The question has been raised as to whether a test should be set up according to the individual's threshold for certain flavors. For example, we had a man who was consistent in his judging, with a correlation

factor of 0.9, but he failed on the triangle test because he was unable to detect "buttery" flavor. The low concentration was below his individual threshold. We are going to consider this problem in planning our next training period.

J. C. HENING: Training with a specific product is very important. This is well illustrated with dairy products on which flavor evaluation has been taught in the schools and colleges for over 30 years. A good milk judge can readily detect oxidized, rancid, bitter flavors, and flavors caused by bacterial contamination. This flavor evaluation of dairy products has helped to promote research to overcome those off-flavors and I think is responsible in part for the high quality of milk, ice cream, and other dairy products which we enjoy today.

MILDRED BOGGS: It might be of interest to note that we found that the time required for training for different characteristics varies. With carrots, for example, it took 5 weeks for flavor, 3 days for tenderness, and 1 week for color.

Committee Report

The committee emphasized the point that results are only as reliable as their source. A sound training program increases the reliability of results. The experiment should be designed so that the person responsible for the organoleptic evaluation can have the judges trained when the product is ready for testing.

Points to consider in training judges

(1) Panel members are preferably trained on the products to be tested, although acuity tests are useful for specific problems.
(2) Trainees should be given an understanding of the problem and what is expected of them.
(3) Interest is stimulated by having judges participate in setting up the score card and thoroughly understanding its use.
(4) Simplicity in the score card is important during the training period.
(5) Memory association is important and persons can be trained to taste with this factor as an aid.
(6) Whenever possible, there should be a reference sample — the food product itself, wax models, etc. — and agreement of opinion concerning it.
(7) It is desirable to have samples of food to illustrate various components and the degree of each component.
(8) Upon completion of training, the panel trainees should have an opportunity to compare their scores with those of experienced panels.
(9) The reporting of results on completed projects increases interest in the training program.
(10) A judge unable to evaluate one product should not be labeled as incapable of judging another without adequate testing.
(11) The training and evaluation of the judges' performance should be a continuous process and retraining may be desirable when judges' performance falls out of line.

Criteria used to evaluate adequacy of training

Training was considered adequate when:

(1) Scores fell within allowable limits of the trained panel members' scores.
(2) Individuals scored the same for four consecutive trials.
(3) All panel members' scores agreed in the judgment of the specific components under consideration.
(4) Individual and panel performance met certain reliability limits as set up by statistical methods.

Need for further investigations

Further research is needed to discover how much, if any, carry-over of "know-how" takes place in judging food products other than the one on which judges are trained.

COMMITTEE: Vera Brastow, J. C. Hening, Cora Miller, Helen A. Moser, Chairman.

Methods of Checking Performance of Panel Members

Discussion

MARY L. GREENWOOD: We found in judging potatoes treated with benzene hexachloride insecticide that the quality of the judging varied markedly from day to day. We used two treated and two untreated samples, and when we were testing some of the antidotes for the benzene hexachloride, there were also duplicates of those. Should we say that they have to judge all four samples right, or three out of four? At what point should they be ruled out?

MILDRED BOGGS: We do a good deal of work along this line, checking performance by any method we can work out that is adapted to the data and the features of performance of interest. In my vegetable section, in general, we do not discard a taster after a test is completed unless we have enough cumulative evidence over a relatively long period so that we decide to drop him from the panel and never use him again. Otherwise we include all data and carry enough judges to absorb a few bad data.

We would like to have methods so we could cumulate a record of performance of all individuals on all types of tests, but in some cases this is difficult because some sets of samples are more difficult than others and comparative performance is not simple to cumulate.

We do have cumulated data from results of a year and a half's work on peas with the triangle test (exhibit 10). First of all we classified all tests into 3 groups: The most difficult in which only 39 percent or less of entire panel identified duplicates; medium difficult in which 40 to 60 percent identified duplicates; and the easiest group, 61 percent or more identified duplicates. Then within each group we compared each judge with the panel by dividing the percent correct of each individual by the percent correct of the entire panel. The entire panel in this calculation is actually the first 14 persons of the exhibit. The others for one reason or another were on the panel for only a short period. You can see considerable variation in individuals here but all of these 14 persons (our present panel) are better than others we have tried out for the panel.

EXHIBIT 10. METHODS OF CHECKING PERFORMANCE OF JUDGES
Triangle test data—frozen peas (May 25, 1948 to Jan. 5, 1950)

Judge	Samples for which panel = 39 percent or less correct identification*		Samples for which panel = 40–60 percent correct identification*		Samples for which panel = 61 percent or more correct identification*	
	Number of tests	Individual, percent correct / Panel, percent correct	Number of tests	Individual, percent correct / Panel, percent correct	Number of tests	Individual, percent correct / Panel, percent correct
Bean	65	0.8	78	1.1	52	1.1
Collings	51	1.3	96	1.1	55	1.2
Cushman	51	1.1	61	1.0	44	1.2
Hanson	56	1.0	78	1.0	44	1.1
Harrington	55	.6	80	.8	53	.9
Hendel	49	.6	72	.6	47	.9
Lindquist	40	1.1	63	1.2	44	.9
Olson	66	.6	76	1.2	53	.8
Roberts	40	.9	62	.9	39	1.0
Seamans	67	1.0	84	1.0	58	.9
Simone	52	1.1	92	1.0	55	1.9
Stanley	36	1.3	64	.8	44	.8
Uhvits	54	1.8	71	1.3	38	1.1
Wolford	55	.8	69	1.0	44	.9
Bittner	20	1.6	34	1.3	23	1.2
Dietrich	13	1.0	28	.5	13	.3
Lee	46	1.1	22	1.1	30	.8
Taylor	14	.9	22	1.4	11	1.2
Nimmo	18	1.2	29	.8	14	.7
Nobles	8	1.6	6	2.0	12	1.1
Witebsky	44	1.0	50	1.0	29	1.0

*The panel used here is the first 14 judges.

Exhibit 11 shows another method of checking performance. It is used when you have duplicate scores for, say, 20 or more samples. The samples here were scrambled dried eggs and the panel was in operation for about 2½ years. Whenever we accumulated duplicate scores for about 40 samples for any judge we calculated the correlation coefficient and the error. If a judge was absent when extreme samples were scored, we waited until he had some extreme scores. Thus the samples for each period in exhibit 11 are not exactly the same, but they always cover 40 or a few more samples and include approximately the same range of qualities. The exhibit does not indicate any general trend for improved performance with longer experience, although we, of course, trained judges pretty well before using them on the panel. The exhibit also shows that good judges sometimes had bad periods even though each period covered a fairly long time.

Scoring tests of the type, four samples per day with four replications, give us the most trouble. We do use analysis of variance on each individual's scores but this has shortcomings. For example, if most judges

EXHIBIT 11. METHODS OF CHECKING PERFORMANCE OF JUDGES

Correlation coefficient, 1st vs. 2d tasting—Dried eggs (Aug. 23, 1943 to July 1, 1945)

Scores analyzed	All periods		Correlation coefficient						
	Number of pairs	r	1st period	2d period	3d period	4th period	5th period	6th period	7th period
Snell	394	0.87	0.88	0.80	0.87	0.81	0.80	0.80	0.98
Boggs	358	.85	.86	.82	.80	.88	.81	.82	.86
Fevold	297	.85	-------	.80	.78	.85	.77	.81	.99
Smith	149	.84	-------	-------	.83	.85	.83	R	-------
Lausten	194	.78	-------	-------	-------	-------	.75	.81	.77
Morris	406	.78	.83	.80	.72	.72	.74	.83	.69
Lewis	341	.78	.85	.76	.73	.83	.81	.73	.79
Rost	192	.78	.77	.85	.78	.68	R	-------	-------
Michener	365	.77	.81	.80	.68	.73	.72	.80	.67
Kester	179	.77	-------	-------	-------	-------	.77	.78	.75
MacDonnell	110	.77	-------	-------	-------	-------	-------	.70	.84
Brandon	70	.76	-------	-------	-------	-------	-------	-------	.76
Stark	297	.76	.73	.77	.83	.74	I	.57	.87
Hirschmann	307	.75	.82	.73	.75	.69	I	.65	.82
Dimick	169	.75	.78	.59	D	-------	-------	-------	.77
Reeve	234	.73	.75	.84	.71	.63	I	-------	.74
Prater	124	.73	.80	.56	D	-------	-------	-------	-------
Klose	104	.69	.79	.54	D	-------	-------	-------	-------
Boyle	66	.69	-------	-------	-------	-------	-------	-------	.69
Vollmer	147	.68	-------	-------	-------	.59	.73	.68	R
Graham	162	.67	-------	-------	.73	.67	.65	R	-------
Larsen	73	.66	-------	-------	-------	-------	-------	-------	.66
Binkley	95	.62	-------	-------	-------	-------	-------	.66	.59
Wolford	69	.61	-------	-------	.61	D	-------	-------	-------
Nielsen	71	.61	-------	-------	-------	-------	.61	D	-------
Jang	49	.60	-------	-------	-------	-------	.60	D	-------
Fox	83	.59	-------	-------	.62	.55	D	-------	-------
Shuffer	89	.55	-------	-------	-------	-------	-------	.43	.64
Dutton	193	.54	-------	.51	.47	.52	.58	D	-------
Edwards	114	.51	-------	.58	.41	D	-------	-------	-------
Pool	73	.50	-------	-------	-------	-------	-------	-------	.50
Debeau	93	.48	-------	-------	-------	-------	-------	.41	.56
Wilbur	54	.39	-------	-------	-------	-------	.39	D	-------
Panel averages		.96	.91	.96	.97	.95	.97	.96	.95

R = Judge resigned; I = Judge absent; D = Judge dropped from panel.

score four samples 10, 9, 8, and 7 and one judge scores them 10, 10, 10, and 1, this one judge gets a high F. Is he a good judge? I don't think so. Suppose also that we have trained intensively so we should be agreed on what is presence and absence of rancidity, but one judge scores samples in the opposite direction from the others. His ratio is also misleading.

We have used various numerators, other than the one in the F-ratio, not squaring differences and considering direction of scores, but do not have just what we want yet. Also we have no way to combine tests to give a cumulative figure. All these things can be picked up with examina-

tion of scores, but we haven't found a routine cumulative system yet. In case someone is interested in working on this, I think I should mention that we are interested in magnitude of differences and error of differences, not variation of the scores themselves.

We have notebooks full of data from which we might learn some things about individual's performance, but not enough time to develop methods to analyze them. Also we should set up some of our new experiments, the main purpose of which is effect of treatment, in such a way that we could learn more about individuals, but there are too many things that need to be done to get the very best plans developed.

L. C. CARTWRIGHT: First, I want to say "Amen" to what Miss Boggs just said. That is a very serious problem. I realize many of us fail to appreciate enough the value of statistics. But I want to say that an error which is often made by workers is trying to depend too much on statistical treatment and trying to get some significant interpretation out of experiments which were not properly planned in the first place.

In checking performance of judges, we depend on relatively simple statistical methods, simply because we have to do so. We keep a running check on each panel member, observing sometimes without making calculations. We do check up from time to time on individual panel members to see if they are deviating too much from the panel average. If a judge deviates in some particular session, we do not hesitate to throw out his results for that one session, and still retain him on the panel. Maybe he was rushed, not feeling well, worried, or for some other reason was "off the beam" on that particular day. Our experience leads us to consider this sound practice.

QUESTION: Aren't your panel members encouraged not to come to the session at all under those circumstances?

L. C. CARTWRIGHT: Yes, but they don't always realize themselves that their judging ability may be off. As an example, say we are scoring five unknowns in a given session with a couple of coded controls. If one member deviates on the total score on the basis of 100 by more than 20 points from the panel average just on one of the samples, even if he hits the panel average exactly on the rest of the samples, we throw out all of his results.

If a consistently good judge scores out of line at a session, there is the possibility that he was confused, so we may ask him to rescore. If the rescoring falls in line with the panel, we use that and throw out the original scores. We then call him in and discuss the two sets of scores with him. This helps him avoid a similar mistake in the future. Sometimes we have the panel member rescore more than once a particular factor of a sample on which his scores deviate from the average. If his scores are consistent in their deviation, we include his score in the panel.

QUESTION: Do you ever find that some panel member or group of panel members tends to develop unconsciously certain attitudes or "fixes" toward one characteristic or another?

L. C. CARTWRIGHT: Well, I don't believe we have isolated that as a cause for a trend in deviation. It may have been one of the factors. We find that we are apt to get a scattering of scores without reproducibility when we don't watch carefully the distribution of the quality of samples.

There is always the problem that a particular sample may be scored

lower with a group of samples of higher quality than it would be with a group lower in quality than the sample in question.

QUESTION: Do you explain the objectives of the study to your judges?

L. C. CARTWRIGHT: We generally make a point of familiarizing the judges with the general problem and what we are going to do. That builds up interest in the judges, which is essential to get the best results from their judging.

HELEN MOSER: We had an interesting instance in which one man on our panel continually indicated a metallic flavor in oil. He was the only person to do so, and ordinarily one person's response is not too important. We recently analyzed the metal content of oil over a period of 6 years and found that the samples varied greatly in their iron content. The samples which this man had been judging were found to have a high iron content. In the concentration studies which we have been running on copper and iron, we find that these people vary greatly in their thresholds for the recognition of iron. This one person could detect it at very low levels of concentration.

COMMENT: We have found that taste factors from an individual's eating habits are carried over into the testing and often account for deviations in results. It is usually difficult to change these habits of taste.

Committee Report

Checking performance of judges is necessary because even the best judges will vary in the quality of their work. Checking should be frequent, preferably every day.

A judge's performance may be checked against himself for consistency through use of (1) a coded control incorporated in each test set of samples, (2) correlation coefficient between first and second tastings, (3) triangle test, (4) identification of duplicates, both treated and untreated, which are included in the test series, and (5) analysis of variance.

Performance of the judge may be compared with that of the group by measuring his deviation from the group mean on test samples.

Following are some practices used in regard to eliminating judges and judges' scores:

(1) A judge's scores for the day are eliminated when he misses the standard control by more than 5 percent.

(2) If a judge deviates by more than 20 percent from the panel average on a test sample, he is either asked to retest the whole series for the day or his day's scores are eliminated.

(3) Scores are not eliminated except on the basis of an accumulative record of at least 2 months which justifies his permanent elimination from the panel.

(4) Judges are eliminated in identification of duplicate samples if they identify fewer than three out of four samples.

Further investigations are needed to determine criteria on which to base elimination of judges.

COMMITTEE: Claude H. Hills, Jessie C. Lamb, Bernadine Meyer, Chairman.

Discussion

GLADYS E. VAIL: We tested sausage with different seasonings for flavor changes during storage. We encountered differences in texture as well as in seasonings. In order to get uniform samples, the sausage was cut into 1-inch slices, heated 30 minutes at 400° F., cooled for 15 minutes, and then a 100-gm. sample with the addition of 75 ml. of water was blended for 3 minutes in the Waring blendor. After blending, the sausage was placed over boiling water for 15 minutes to get a good blending of the seasoning with the meat. It was found that the flavor dropped faster than the aroma during storage. Oxidation was also measured and found to be increased by storage.

MARY L. GREENWOOD: One of the problems in our potato testing was to maintain the texture in the samples as nearly like the original as possible. We tried various methods, including quartering the tubers lengthwise, putting them together again, wrapping in aluminum foil, and baking. The foil was first punctured to allow for the escape of steam and to prevent sogginess. When the potato was removed from the foil, the samples were scattered among the judges so that each got no more than one cube from a given quarter of a potato. This method seemed to give the most reproducible results.

MILDRED BOGGS: We have had difficulty in scoring peas because when the flavor goes off, so does color and texture. We can puree the peas and add vegetable dye or set samples under special lights to mask color differences, but even the Waring blendor will not eliminate differences in texture. We are at present comparing the scores that we get for the whole peas, for the cooking liquor that is drained off the peas, and for a supernatant prepared by blending the peas in the Waring blendor and centrifuging. As you see from the results in exhibit 12, we are getting similar scores when the peas, the liquor, or the supernatant are judged, and we are most encouraged. Actually all data on this project so far indicate that either differences between samples are larger or error smaller with supernatants than either of the other forms. We feel, however, that we must check this with all the off-flavors of peas that we know about before we can accept this form of peas for routine tests.

EXHIBIT 12. SCORES FOR FROZEN PEAS SCORED IN THREE FORMS
(4 replications — 10 judges per replication)

Form	Average off-flavor scores			Least difference at 1-percent level
	No delay	Delay 4 hrs. at 90° F.	Delay 3 hrs. at 90° F. plus 6 hrs. at 80° F.	
Whole peas	5.0	3.6	2.6	0.6
Supernatant [a]	5.0	3.6	2.4	.3
Cooking liquor	5.0	3.2	2.7	.6

[a] Supernatant obtained by blending cooked peas and their liquor in a Waring blendor, centrifuging, and pouring off liquor.

In another study in progress now we are trying to determine whether the kind of off-flavor that develops under three treatments with frozen peas are alike or different. The treatments are delay between vining and blanching, high temperature (+10° F.) storage of adequately blanched undelayed peas, and underblanching followed by −10° F. storage. We find it very difficult to identify different kinds of off-flavor in a product like peas which have a fairly pronounced natural flavor, but we are working on this phase of the job.

In still another study on frozen peas we are trying to develop a reference standard for routine taste tests that does not change in flavor during about 1 year of storage. For this purpose we store peas handled under optimum conditions at −70°, −30°, and −10° F., and compare these at various storage periods. We also compare the −70° and −30° samples stored 1 year with freshly harvested material of similar history. The −10° peas are definitely not suitable for reference material, but both −30° and −70° look promising. We of course need to develop reference standards for many commodities but this is a start.

HELEN J. PURINTON: I mentioned yesterday our work with frozen strawberries. The purpose of the work was to evaluate new varieties along with standard varieties grown in our region and to evaluate their adaptability to preservation by freezing. The flavor of the frozen berry after storage and the effect on flavor of different methods of preparation of the frozen sample were both evaluated in the palatability tests. We used dry sugar, sirups of varying sweetness, and sliced berries as well as whole berries in preparing the samples. We wanted the berries judged for flavor alone and ran into discrimination among the judges against the sliced berries and the unsweetened samples. So we blended them in the Waring blendor, added given amounts of sugar or sirup, and scored them in pureed form. The same panel of judges could not perceive any difference in the blended samples between the flavor of berries sweetened before freezing and those to which sugar had been added just before blending. We therefore decided that the first results, in which the judges gave low scores to unsweetened frozen berries, were not valid.

Graduate horticulture students are given preliminary training before going around the country on panel judging teams. First, they are put through the primary taste tests described this morning. Then, supposing we are judging apples, they will test some standard varieties, with open discussions on the characteristics and method of evaluating. The students will then get some practice judging perhaps a new variety or a variety which is being studied for adaptation to our climate. We usually offer four samples, with a standard reference sample, perhaps a fresh McIntosh. They will judge the variety of apple as fresh, in baked form, as applesauce in canned and in frozen form, evaluating for flavor and color. We consider a variety good for our particular region when the judges evaluate it as acceptable in these four different forms. Needed research is indicated through this judging work when some characteristic of an apple is not pleasant to the judges.

MILDRED BOGGS: We are all confronted with this problem of varieties. We don't know the answer.

ERNEST C. CROCKER: The season affects the characteristics, too. Baldwin apples are not very good in the fall but are tops about December or January. Other varieties, like the Winesap, are used all winter. Now

are these to be judged several times during the season so as to get the fall-off in quality or are they to be judged just once?

FRANCES O. VAN DUYNE: We found that a judge would consistently rate a particular variety high and then as the flavor of that variety went down during the season, her scores went down. Another judge would rate another variety higher. We didn't know anything to do except to average all the scores and hope that that represented the people's preferences. We didn't feel that we could select one variety and say that the flavor was superior.

MARY L. GREENWOOD: How did you take care of the changes in the apples as the judging proceeded?

FRANCES O. VAN DUYNE: We tried to select representative samples with comparable stages of maturity and we offered the judges more than one slice.

QUESTION: Were the slices taken from just one apple?

FRANCES O. VAN DUYNE: No. We used parts of different apples.

COMMENT: We try to limit the kind of information we want from the judges by asking them to score for only one characteristic.

A. KRAMER: I can't say much about apples, but with some vegetables we find that we can group varieties and treat them as a group. With snap beans we find that we can group them roughly into three classes. The same thing can be done with peas. The early varieties, the intermediate, and the later peas can be treated as three different units.

COMMENT: We find that with limas, too. The Henderson type and the Fordhook type can be considered as two separate groups. The two groups cannot be compared.

Committee Report

Reproducibility of cooking and serving procedures

Different products present specific sampling problems. In testing meats, it is recommended that sides of animals be matched. Light and dark meat of poultry should be judged separately. Similar cooking time and temperature and the same kinds and amounts of added ingredients must be used for samples which are to be compared. Apples for judging should come from the same tree. Spice tests are most accurate with plain white sauces.

Similar containers should be used for serving each time. These should be plain and of material which does not affect the flavor of the food product.

Size of sample

Several factors affect the size of samples. Smaller samples of strong-flavored foods can be used as compared with bland-flavored products. Kale, for example, fatigues the sense of taste when samples are too large or too numerous. The total amount of material available for testing and the number of replications sometimes limits sample size; for example, there may be only a small quantity of canned chicken of a specific pack.

The amount in one unit which must be tested by all members of the panel frequently determines the size of the sample. To illustrate, a potato tuber may be quartered, and each quarter cut into thin sample slices which are fitted back into place before wrapping in aluminum foil, puncturing the foil, and baking. Samples should be large enough for adequate tasting but not so large that they overstimulate the taste buds.

Temperature of food

All samples of one food should be tested at the same temperature. Most foods should be tested at the temperatures at which they are normally served. There are some exceptions to this rule, mainly for foods that are ordinarily eaten very hot or very cold. Room temperature is satisfactory for many products. Flavor differences in frozen strawberries are more apparent when they are completely thawed. For that reason, although the berries might normally be served before they are completely thawed, analytical rating of flavor should be carried out with thawed fruit.

Color differences

When flavor is being rated, it is desirable to eliminate the factor of color differences in the samples. This may be accomplished by using special lights, dyeing, or by having judges trained to taste without looking directly at the sample. The successful use of sodium vapor lamps was reported; they mask approximately 50 percent of the color.

Sampling problems which need further study

Further studies are needed in regard to size of sample, method of selection, and physical and chemical techniques to assist in achieving uniformity. The techniques must be developed for each product and in terms of the questions to be answered. Methods of correlating results need to be developed in order to set up general principles of sampling.

Problems are especially acute in the sampling of fresh, canned, and frozen foods. Blending into one mass can be used to obtain uniformity of sample when testing for flavor without considering texture and color. Grinding or blending has been successfully used for testing flavor of strawberries and sausage and rancidity in chicken and peas. In tasting the peas, the supernatant fluid from the blended, centrifuged peas, as well as the cooking liquor, was used for the sample. Potato samples have been prepared by using six to eight tubers as a representative sample and cutting the potatoes in quarters or eighths as described above.

It is difficult to maintain standard reference samples even in frozen storage. A temperature as low as $-70°$ C. may achieve better results than higher temperatures.

COMMITTEE: Verz Goddard, Georgia Schlosser, Frances Van Duyne, Pauline Paul, Chairman.

Conditions of Judging and Judging Room

Discussion

L. C. CARTWRIGHT: It is most important, of course, that all samples be coded and that they be similar in appearance, size, and color when they

are being evaluated for flavor or odor. They should be served in similar containers of the type in which the food is ordinarily served. Panel members should be trained to taste a similar amount of each sample. We do not specify a definite amount for all panel members. We believe that different panel members will use different amounts of the same food to get accurate evaluations. The panel member should be permitted to decide what amount he needs in order to evaluate the food.

The temperature of the food sample should be close to that at which the food is normally served. It is desirable, however, to avoid extremes of hot and cold. We believe that coffee and soup, for example, should not be tasted when very hot because accurate distinctions between flavor and odor are more difficult when the product is tasted quite hot. We have compared scores of such samples tasted hot and at a lower temperature and find that there is a difference in the scoring.

In preparing samples, we use either home or commercial preparation, depending upon the problem. We blend and puree many foods in order to eliminate the factors of color and appearance when evaluating odor, flavor, and aftertaste. We have found it helpful to train panel members to look at the samples only enough to see what they are doing but not to look directly at them while they are tasting. This helps to eliminate the factor of appearance. We have also used colored lights, masking, etc. In evaluating spices, we prepare a food with a smooth profile — according to the Cairncross and Sjöström system of flavor profiles (Food Technol. 4: 309. Aug. 1950) — for comparing with the same food in which the spice to be tested stands out in the profile. We have used successfully plain white sauces for testing spice flavors.

MILDRED BOGGS: We have used sodium vapor lamps with such success that we are installing them in our booths. They are inexpensive and effective with about 50 percent of the items we test. This procedure was initiated by Dr. C. C. Nimms of our laboratory.

L. C. CARTWRIGHT: In evaluating flavor, we caution our judges against too frequent tasting of the standard control. Often, an experienced panel member may find it unnecessary to taste the standard control at each session.

In connection with judging room conditions, we think that some investigators have placed too much emphasis on some factors there. We feel that the room should be comfortable without distracting noise or odors. Our panel members do not score in the presence of each other, except during training periods. It seems unnecessary to refine one phase of the experiment (such as judging room conditions) too far beyond what the maximum obtainable accuracy due to other factors will justify.

MILDRED BOGGS: I will describe an experiment we carried out on sulfite testing. During the war, we were scoring dehydrated foods, many of which have sulfite in them. The experimental errors were large, which we believed was due to carry-over of the sulfite flavor from foods containing it to foods which did not have it. This seemed to happen when a food with sulfite in it was tasted just before one without it. We also thought that after tasting one sulfite, the acuity was dulled for the next sulfite.

In order to test these theories, we set up a test using mashed potatoes to which sodium bisulfite solution was added after cooking but before mashing. We tasted at one session a control and two sulfites of the same concentration and at another period two controls and one sulfited sample. We set up different concentrations — using 12, 25, 50, and 100 parts per

million — but only one concentration was tasted on one day. We had our judges taste a labeled unsulfited control, then the first unknown, then the labeled control again, and the second unknown, and so on. Judges rested ½ minute between unknowns. They were not permitted to retaste any sample and, of course, scored the samples in the prescribed order.

The results of this study showed that under the conditions of the experiment sulfite flavor was not carried over from a sample containing sulfite to one that did not, but one sulfited sample dulled acuity of the taster for a second sulfited sample of the same concentration (exhibit 13). The second sulfited sample with 100 p. p. m. tasted about like the first sample with 25 p. p. m.

EXHIBIT 13. INFLUENCE OF THE FIRST SULFITED SAMPLE TASTED ON SCORE FOR THE SECOND SULFITED SAMPLE TASTED — MASHED POTATOES

(The 2 sulfited samples of one test contain the same amount of SO_2)

Test No.	Concentration of sulfite (p. p. m.)	Average score[a]			Least difference at 1-percent level
		No sulfite	First sulfited sample	Second sulfited sample	
1	12	6.9	5.7	6.2	0.8
2	25	6.8	4.4	6.1	.8
3	50	6.8	3.6	5.5	.5
4	100	7.0	2.8	4.9	.6
		First control	Second control	Sulfited sample	
5	50	6.7	6.8	3.7	.4
6	100	7.0	6.6	2.8	.8

[a] Score 7 = no sulfite flavor; score 1 = much sulfite flavor.

I believe that effects like this enter into all of our tasting work, but that the effect is accentuated with sulfite and no doubt certain other substances.

HELEN MOSER: I would like to show the plans of our judging room as they illustrate what can be done on a small scale. The only room available was an inside storage room 11 by 16 feet, so we planned the room to meet our needs. There are no outside windows and so we put in an air-conditioning system which maintains a temperature of 78° F. and 40-percent humidity. The panel member comes in from the corridor and sits down in one of the four booths. A dual light system indicates his presence and the samples are then offered from the preparation area through a sliding door at the back of the booth. There is no contact between the person in the preparation area and the panel member. The room is odor free. (Slide shown of judging booths and preparation area.)

The samples of oil have been warmed on the hot plate which is arranged to heat samples to the same temperature each time. The panel member tastes the sample and leaves through the office area where he sees what he

has tasted and compares his results with those of the others. This opportunity for comparing his scores helps to maintain interest in the judging. We also bribe the judges with cookies at this period.

Committee Report

The following points appear to be agreed upon and important:

(1) Samples should be coded.
(2) Judges should not judge in the presence of each other; for example, they should judge at different times or in individual booths.
(3) Judges should have no contact with preparation of samples.
(4) Conditions of serving should be near normal or neutral with respect to temperature of sample, utensils, seasonings.

A desirable time for judging is before meals; around 11:00 a. m. is frequently used. The number of samples tested at a session and the time interval between sessions depends upon intensity of odor and flavor and the experience of the judges.

The character of the sample determines whether a mouth rinse is used and the type of rinse desirable. The investigator should watch for dulling of acuity of judges when testing several samples of certain types, for example, sulfite flavor.

Two points of view were represented on conditions of judging room:

(1) Rigid control of such conditions as temperature, humidity, light, color, and odor.
(2) Less emphasis on rigid control on the assumption that conditions of the judging room introduce less variability than other conditions not so well defined.

Area for further research

Information is needed on the relative precision of judgment under rigidly controlled conditions and under controlled but less rigid conditions as a guide to laboratories where circumstances prevent rigid control.

COMMITTEE: Barbara McLaren, Grace Schopmeyer, Alice Briant, Chairman.

Summary of Factors Determining Accuracy of Tests

Discussion

ERNEST C. CROCKER: The sense of smell brought me into food work in the first place. I was a chemist interested in organic chemical odors. Odors impressed me then, and still do. As I approach food today, the odor is what speaks first and seems most important. Odor is the most characteristic thing — not color, not taste, but odor.

In making a detailed examination, you first bring the food to the nose for odor. You can get a little bit of taste sometimes with a product like oil of cinnamon or oil of anise because you breathe in enough so that it dissolves in the saliva and you can taste the sweetness, but what you get from your nose is nearly all pure odor.

When a product is placed on the tongue, it is first tasted, then later odor registers, for which you usually make allowance, having smelled it

before the tasting. Then there are factors of feeling, examples of which are astringency and the metallic tastes, which are very important, in addition to coolness, warmth, etc. About 10 to 15 seconds or a minute later, you will note the aftertaste, which determines whether you will ever bother to taste that article again. An unpleasant aftertaste — as when rancidity shows up — will prejudice you against the food.

Odor is useful for identification. To those who are odor-conscious, identification by odor is far more important than it is by any of the other senses. There is so much to odor. There are hundreds of thousands of different shades of odor and one can determine with a great deal of precision what the chemical is and the amount, in many compounds. Then there is the appraisal of quality. The odor tells you what condition the product is in; it gives a pretty good idea of bacteriological action and what thermal abuse the food has been put through. Odor tells you also whether or not you are going to like the food.

Some people believe that appearance is the first consideration in vegetables. I would go back farther to another factor — and say that toughness is the first consideration, for without that, the product will not stand shipping and will likely never get into your home at all. Suppose it is tough enough to stand shipment and looks well enough to entice people into buying it — then there is the question of whether or not they will eat it.

Have you ever been in a grocery store when they cut open a nice Hubbard squash that scents the whole store? Well, of course, you are going to go over and buy some of that squash, that is if your nose is activated. You'll take it home and enjoy eating it. There is another squash that is cut up, but it smells as flat and dead as can be — you just let them keep that one.

You can become odor conscious by training — it only requires a simple little operation known as the sniff. You sniff the odor to the smelling area in the nose. Normally, the air currents go through various channels and fail to reach an appreciable part of the smelling area, so that you are not commonly aware of what odors there are in the air around you.

How many in this group here were aware that one of the waitresses at the dinner last evening was heavily scented with perfume? And what perfume was it? I doubt if more than three or four noted it and if more than two or three could tell what kind it was. Two of us sitting side by side spotted it instantly and recognized the kind of perfume. To develop odor consciousness like that, you've got to be forever taking little sniffs, which lift the air into the smelling area of the nose. It calls for effort at first, but after a little while, you don't notice that you are doing it and a whole new world opens to you. Some people say that they are not much interested in a dog's world or life, but you can get a great deal more out of it than you might think. It really is worth cultivating!

You can become nose-conscious and nose-directed, and it is a great aid. In our work in our flavor laboratories, we insist on odor training of all the people. We don't put them all through odor classification in testing measurements and all that sort of thing, but we do make them conscious that there is such a thing as odor and teach them to smell first. We devote fully half of the flavor panel session to smelling. Of course, with roasted coffee or coffee beverage, the odor is the primary thing. With products that require only odor evaluation you can handle as many as a dozen or two samples at one time. You are not limited to six or eight as you are in tasting. You can test a great many samples without fatigue. Also, you are not loading yourself with food.

The elements of odor, like those of taste, need not all be pleasant. Every characteristic odor has pleasant notes, and in some instances, unpleasant ones. There are often little kick notes that are quite unpleasant. For instance, in freshly roasted coffee, there is a skunky note that belongs there. This skunky note is a desirable thing, but skunky smell in beer is most undesirable. This skunky odor compares with bitterness or sourness in taste, which may be desirable in small amounts to do exactly what's needed to complete a flavor. Likewise, very disagreeable odor sensations are sometimes needed to round out the total effect.

The "flavor profile" is an aid for describing flavor. Using the hand to illustrate, one can say that the components that have been blended so that nothing sticks out are represented by the palm of the hand and all the notes that stick out beyond that are represented by the fingers. You may have a product that is highly seasoned so that several "fingers" stick out. Perhaps you want to make these points less conspicuous. You can get it so by putting in more ingredients or bring up the general average so that the fingers are less prominent. You generally want a few characteristic notes to stand out in any final flavor, but you also want a great body of blend with all notes fairly well submerged. It is difficult to discover the individual components in a product that is beautifully blended. On the other hand, a product with "fingers" of flavor sticking out can be analyzed easily.

In our laboratory testing, we do not use elaborate gadgets. Our primary considerations are comfortable seating and a homelike atmosphere for the panel judges. You sit down at the table with two or three others and make believe you are enjoying yourself — in fact, you are for a while. You are served a helping of food and given a piece of paper and a pencil. You start by smelling the food. Of course you wouldn't do that at home, or at any rate be caught doing it! After smelling the food, you taste it, and of course you later experience the aftertaste. While the food is still there, you discuss it.

With a new product, you start with 3 or 4 people and work out some terms of evaluation, on the basis of the characteristics that are of greatest importance — astringency, bitterness, or others as the case may be. Later a panel of 6, 8, or 10 will evaluate the product on a numerical scale. Thus, the first examinations of the product are really like a survey party sent out to discover what work is needed to be accomplished. The additional personnel of the expanded panel then follows the path laid out by the survey party. This method works well with organic chemicals as well as with food, and enables one to swing quickly from one product to another.

To get back to odor, there are the terms "whiffing" and "sniffing." When a product has a strong odor, you hold it a little distance from you and fan the odor toward you — that is whiffing. If you draw it into the smelling area of the nose, that is sniffing. There is an optimum quantity that you need for a sniff. Your nose will take only a certain amount. If the odor is weak, you may strain to smell it and still the result will be unsatisfactory. It is most satisfactory to have plenty of odor and to help yourself to just as much as you need.

Another chemist and I worked out an odor classification system some 23 years ago which divided odors into four components: Fragrance, acidity, burntness, and caprylicness. Fragrance is the sweetness of flowers, spices, honey, etc.—for example, extreme pleasantness and sweetness. We called this component F. Acidity we called A and exemplified it by the sharpness of vinegar, cheese, and in some cases, of

rancid butter. It does not depend upon the hydrogen ion, although many volatile acids are able to stimulate it to a high degree. Alkalies, however, and foods independent of chemical acidity can also possess a high acidity odor. Then B is burntness. The Greeks used that expression 2,300 years ago. They called it "the odor of burning flesh." The term "caprylic," one which we introduced, means "goaty." The odor of a wet dog is the smell we mean — that is perhaps better understood than the smell of a goat.

These four components — fragrant, acid, burnt, and caprylic — can be recognized in all odors. Fragrance is pleasant to human beings, acidity is usually pleasant in fair amounts, burntness only moderately pleasant in fair amounts, and caprylicness hardly pleasant at all. Caprylic is the characteristic smelled in rancidity, perspiration, and gasoline. In general, we close our noses to it because we find it unacceptable. However, without a little caprylicness in every odor, it is flat. You might call this component the seasoning in all odors. There are instances where we want it in high amounts. An example is limburger cheese, which if liked, is liked strong. We accept it in some instances where we know it is safe, but in general a caprylic note means bacteria at work, and we are alerted by it. When we know we can relax and enjoy it, we can get something pleasant out of even caprylicness. With human beings the general factor of pleasantness goes about like F, A, B, and C. With dogs, it is probably just exactly the opposite.

We discovered that each odor contains some of each of these four components and we set up a system of evaluating each odor on an 8-point scale. Thus, an odor number can be reached for any odor, which will completely describe it. Using whole numbers, you have $8 \times 8 \times 8 \times 8$, which amounts to nearly 4,000 different numbers. Professional perfumers were able to break down each number into three parts. This brings the total number of odors up to around 400,000. (Mathematics is wonderful — you really can do things with figures!) To give an example, acetic acid is about 3 degrees of fragrance, of a possible 8; it is all 8 in sourness, about 1 degree of burntness, and about 3 or 4 degrees in caprylic. When you say 3813, you completely describe the odor of acetic acid. That amounts to a simple description of odor, whether or not it is entirely accurate, gets down to a physiological fact, or is just an assumption. Unless so simplified, odor description is almost hopeless.

You can create and change odors by adding 5 to this one, 1 to that one, etc. but remember that the 1 to 8 scale is a power series. No. 1 in intensity is one unit and when you go up a step, it is about threefold in intensity, and at 3 becomes 9, at 4, 27, and you see it goes up about 1,000 to 1 in 8 steps. You can't add logarithms so you can't just make additions or averages. On the other hand, you can figure what will happen when you do make additions, but have to keep in mind that you are dealing with logarithmic quantities. As nearly as we can estimate, the difference from 7 to 8 or 4 to 5 is about 3 to 1 for the standards we have adopted. We reserved the digit 9 for some synthetic chemist who may make a product outside the range of accepted standards. I am sure that 9 fits the caprylic and also burntness of mercaptans. This is a rapid word picture of the evaluation of odors.

Then there is the question of diluting-out odors. There are many complications here, including the logarithmic scale, as with light and sound intensities. An essential oil may be of a 100-percent concentration, or 10, 1, 1/10, or 1/1000 percent. What happens when you try to dilute it with something else? Let's take the example of oil of wintergreen being

diluted with an inodorous inert oil. It happens that the odor concentration stays about the same on dilution up to a certain point and then begins to fall off. It makes a straight line on a logarithmic scale and at about 10,000 to 1 further dilution disappears. In other words, there is a range of concentrations where there is more odor than you need for a smell, so that you help yourself to as much as you need, but after a certain point, there is weakening and you have to pull for it, and finally it disappears.

Let me close by suggesting that you consciously study odors for a while and learn the art of sniffing. It will increase your life's interests.

I. D. JONES: As already pointed out by Dr. Greenwood and others at this conference, there are frequently sources of variation other than in the techniques of evaluation of the food products. The information that I will present is the summary of some work done by horticulturists, statisticians, and entomologists at the North Carolina Agricultural Experiment Station on the influences of a number of insecticide treatments on the flavor of peaches.

We used six different spray treatments. Some of the spray materials were known to impart definite off-flavors; others were suspected of doing so. Our standard of comparison was lead arsenate. The trees in plots separated by double guard rows were sprayed with the insecticide to be tested. At harvesttime, fruit was selected for uniformity of maturity, and again in the laboratory at the time of preparation for taste testing.

It was assumed that one person could not effectively taste more than five samples at one time, so the experiments were set up to taste fruit from only five treatments at a time. Each treatment was tasted by 20 different people in each test. The sample submitted to a judge from a given treatment was made up of small cubes from each of 10 different fruits, the material having been peeled and cut immediately before testing. Every judge was given a tray with a coded sample cup from each of five treatments. The order of tasting was specified; in some cases this was determined at random, in others it was controlled so that the effects of order of tasting could be measured. No selection of tasters was exercised and no training was given them.

Although significant differences were observed between treatments, the comparative acceptability of the fruit varied greatly between treatments, depending upon the date of harvest. (Slides shown.) For example, for fruit harvested July 22, 1948, that sprayed with chlordane was found to be significantly better than that sprayed with lead arsenate, the check treatment. However, for the fruit harvested July 27, 1948, the chlordane-sprayed material was significantly poorer than the lead arsenate-sprayed fruit. Similar reversals in the order of treatment means possessing significant or nearly significant differences at the 0.05 probability level were observed in 1949 studies.

We have no explanation for the curious behavior of the treatment means which have been presented for the fruit harvested at the two different dates for a given season. However, it is believed that there should be a realization that this treatment-by-test interaction may be, at times, very large. Such interaction indicates that the process of refining techniques for discrimination between treatments is relatively inefficient in reducing the real source of variation in cases where considerable variability between similar materials at different dates exists.

QUESTION: What were you scoring, characteristic fruit flavor or presence of the insecticide flavor?

I. D. JONES: We were scoring on the basis of good or bad flavor. We designated the order of tasting but found in this particular study that the order was not important in the detection of off-flavors.

GRACE BENNETT: We do cooperative research at Penn State with members of the agricultural experiment station. We are working with storage problems of frozen foods. I will present a problem we have in our work with ground beef. We judge our samples, put them in storage after different treatments, and then judge them every 3 months for a period of a year. We are using some of our cooperators as judges because of the difficulty of making up a panel. They are interested in different aspects of the problem — some are chemists, others are bacteriologists.

We use a rather large amount of meat each time we test. We judge our samples in two groups of six each, with a 20-minute break between the two periods. We judge at 11 a. m. since this is a convenient time for the judges. We train our panels for a week before they start judging, although we feel that this length of time is inadequate. We have the problem Miss Kirkpatrick mentioned, that our material is changing.

Probably our judges know too much about our samples, because while we are tasting them in Home Economics, they are running free fatty acids and peroxides on them in Biochemistry and counting the bacteria in them in Bacteriology. But that hasn't prevented a very interesting happening. As we were getting well into a year's testing, we began to discover that our judges were finding no differences between samples from the different storage periods. They could find differences between samples treated differently in each storage period but not differences from one storage period to the next.

The problem appeared to be that our judges had no standard against which to score. A partial solution was reached in using one of the treatments as a standard. However, this method presents a problem when the material is changing with increased storage. If the judges fail to grade down the standard sample from storage period to storage period, they do not have a proper reference sample against which to score. Our problem is a source of a standard reference sample for the judges.

CLAUDE H. HILLS: We have had a similar experience in scoring for flavor apple juice which has been stored at different temperatures. The first year we used no standard. We just compared the different treatments and were not satisfied with the results. The next year we compared frozen samples. They can be kept for 2 or 3 years. I would suggest that you freeze your meat samples unless you are trying to compare the effect of frozen storage.

GRACE BENNETT: We are comparing the effects of frozen storage so that is not a solution.

DAVID R. PERYAM: Can you reduce the temperature quite low for holding the sample?

GRACE BENNETT: We have a mechanical difficulty there.

COMMENT: We tried to hold a standard sample of canned dried milk relatively constant by storing it at −20° F. After being stored at that temperature for 5 months, it changed so much that anyone could detect the change.

COMMENT: We got the same sort of thing with ice cream. We stored it at −20° F. and it showed definite changes over a 2-month period. We

have had dried milk stored at room temperature up to 4 years without that much deterioration.

MILDRED BOGGS: We always try to select uniform raw material, but as you know, it is difficult to get uniform vegetable samples and there is still variation in different parts of one unit. When flavor is the factor under consideration, we usually chop, grind, puree, or do something to make samples more nearly homogeneous.

We try to limit the number of quality characteristics scored as much as possible. In the first study with a given commodity and treatment we usually score several characteristics because we do not know what is important, but as soon as possible we limit this to one or two if we can.

I have previously referred to our work to develop one excellent quality reference standard. We want to work on others. We are beginning to think that reference standards in the middle of the scoring range are more useful in many studies than excellent quality ones.

Dr. Helen Hanson has given me permission to mention a problem she encountered which illustrates an interesting point about judges. She was scoring 10 samples in a balanced incomplete block, 4 at a time. She found that a medium-quality sample got a high score when judged with poor-quality samples and a low score with good samples. For example, a sample with score 5 would be scored about 7 if all the other samples that day were quite poor, and it would be scored about 3 if all the other samples were very good. This tendency to score comparatively is very important in selecting experimental plans.

There is one more study I would like to mention here. Its purpose was to determine the influence of two or four samples per taste session on acuity and consistency of judges in scoring flavor of frozen corn. We scored all four samples together and also each sample with each other sample but only one pair per taste session. Our tentative conclusions are that judges were no more acute or consistent with two than with four samples per session and that four samples is the more efficient plan (exhibit 14).

MARY L. GREENWOOD: I will say a few words regarding the number of samples. Our judges don't like to go above 8 samples with potatoes. They preferred not more than 6 with kale. They didn't mind 16 samples of blueberries, where they ranked groups of 4, without assigning scores.

MILDRED BOGGS: I would like to report a study on texture of peas. We thought we could not score 24 samples in one taste session even though we were scoring texture only — not flavor. However, we wanted comparative scores for all 24 samples in order to correlate judges' scores with objective measurements of texture. We thought that judges scored comparatively the samples in front of them at one time so wanted to minimize score variation due to samples scored in a given session.

We had the tenderometer readings as a rough guide to texture so took advantage of this in our plan. We arranged the 24 samples in increasing tenderometer order and scored 12 samples per session. On day one we scored order 1–3–5–7, etc., to 23; on day two, 2–4–6–8, etc.; day three, 1–2–5–6, etc.; day four, 3–4–7–8, etc.; day five, 1–4–5–8, etc.; day six, 2–3–6–7, etc. Each of these sets was scored twice. Thus, approximately the same range of qualities was scored each day and each sample was affected about equally by each other sample, without the large number

EXHIBIT 14. INFLUENCE OF NUMBER OF SAMPLES PER TASTE SESSION IN SCORING OF FROZEN CORN

(6 replications — 8 judges)

A. Differences and errors

Holding periods compared	Score difference		Standard error of difference	
	2 samples per day	4 samples per day	2 samples per day	4 samples per day
0–2 days	−0.1	0.0	0.10	0.18
0–3 days	.6	.3	.17	.24
0–4 days	.6	.7	.31	.20
2–3 days	.1	.3	.06	.18
2–4 days	.6	.7	.14	.18
2–4 days	.6	.7	.14	.18
3–4 days	.3	.4	.14	.14
Average of all pairs	.4	.4	.15	.19

B. Significance level for above pairs with varying number of replications

Number of samples per taste period	Number of taste sessions required	Number of times each sample is scored by all judges	Significance level for following pairs[a]					
			0–2	0–3	0–4	2–3	2–4	3–4
2	6	3			**			
2	12	6		*	***		**	
2	18	9			***		***	**
2	24	12		*	***		***	**
2	30	15		*	***		***	***
2	36	18		*	***	*	***	**
4	1	1						
4	2	2			*		**	**
4	3	3			**		***	**
4	4	4			***		***	**
4	5	5		*	***	*	***	**
4	6	6		*	***	*	***	**

[a] *, **, *** indicate significance at 0.05, 0.01, and 0.001 levels, respectively.

of sessions and material that would be required for a balanced incomplete block. Replicate scores for a given sample were remarkably consistent by this plan.

LYLE CALVIN: Where we have several tests, we should use for our error an interaction term composed of both failure of the judges to agree within

each test, and failure of the mean scores on each test to be the same from test to test. Commonly, only the first portion is used for error. Dr. Jones has previously indicated that he has encountered this disagreement from test to test and that recognition of it should be made. If I might use the blackboard I would like to illustrate what is happening.

Dr. Jones gave an example of a test in which a particular treatment was significantly better than the control; however, when a second test was run a complete reversal of results was obtained, that is, this same treatment was significantly worse than the control. Although this is somewhat extreme, we would expect many times to obtain results which vary from test to test, making us somewhat uncertain as to exactly what we do have. A proper measure of error must then include this interaction of treatments from test to test.

If we compare $V_{\bar{x}}$ (the variance of the mean) in several different cases, we can see more clearly the effect of the second portion of error. The formula for $V_{\bar{x}}$ in this case, is

$$V_{\bar{x}} = \frac{V_w}{rt} + \frac{V_t}{t}$$

where V_w is the error as determined from a single test, V_t is the component of variance due to treatment $\times$ test interaction, r is the number of persons per test, and t is the number of tests. This assumes that we have the same number of tasters in each test, although allowances can be made if this isn't true.

On a series of tests conducted this past year on peaches we obtained values for V_w and V_t as

$$V_w = 1.135$$
$$V_t = .055$$

From the use of the formula, comparisons can be made when r and t vary. For example,

r (number of persons)	t (number of tests)	$V_{\bar{x}}$
42	2	0.041
42	1	.083
42	1	*.027
8	3	.065
8	1	.197
8	1	*.142

* The $V_{\bar{x}}$ which would be obtained if only one test were run. In this case no estimate of V_t is available and $V_{\bar{x}}$ would be determined as

$$V_{\bar{x}} = \frac{V_w}{rt}$$

instead of the correct

$$V_{\bar{x}} = \frac{V_w}{rt} + \frac{V_t}{t}$$

This should give you some idea of how much difference is possible. In some types of experiments $V_t = 0$ and then the error within a single test would be the correct one; however, we should realize that this is probably not usually the case and hence we should be using the treatment $\times$ test interaction.

QUESTION: Would you be willing to generalize when you would use the interaction error and when you use the other?

LYLE CALVIN: The interaction term should always be used if we have more than one test. If we have only one test, then we are forced to

assume that our error for a single test is not too low, although this assumption may be quite unrealistic. In this case we are getting apparently significant results which may not actually exist.

QUESTION: Do we eliminate the day-to-day variation when we present exactly the same four samples today, replicates of them tomorrow and the next day, having four means for each sample?

LYLE CALVIN: You have taken out the day-to-day variation from the variance of the mean, which is quite valid, but not the treatment × day interaction.

COMMENT: That is the reason I was asking. We have been worrying a lot but haven't done anything about it yet.

GERTRUDE COX: I might throw in a few comments which I think are very vital to your whole research program. You have a variation in different judges, and apparently you also have a great deal of variation in that your judges perform slightly differently from day to day. For any one day, your mean has in it that failure of the judges to perform the same on different days and you do not have that variation in your error term. When you are testing a mean and there are variations included in that mean, those variations should also be in the error. We have to be mighty careful in comparing means in deciding what kind of variation is making this mean what it is or making two means different. That variation must be in our error term. When you use a few judges on at least 2 or 3 days, you have an error term made up of a combination of the variation of the judges and the failure of those judges to perform the same on different days. We have considerable evidence that this type of variation is very pronounced in animal experimentation.

COMMENT: We deal with daily means. We usually have four replications and four means for each sample. We never enter the original raw scores of all of the different judges. We thought it showed up in the day-to-day fluctuations of the means.

GERTRUDE COX: You are taking the failure of these means to be the same on successive days as the error term. That's a perfectly good estimate of error if you adjust everything to the basis of means.

Committee Report

Uniformity and quality of food

Several methods have been reported for achieving uniformity in food. In judging flavor in peas and stored frozen strawberries, aroma in frozen sausage, and rancidity in canned chicken, the Waring blendor has been used. Blending produces a homogeneous mass which makes for uniformity in the sample. Research is needed to determine the effect, if any, of beating air into the food on the flavor. If a blended sample is used for testing flavor, other samples are then necessary for judging other characteristics such as texture.

Many foods in the natural state are not uniform; striking examples are broccoli with its buds, stem, and leaves, and meats with their several muscles in one cut. If one muscle only is cooked, the effect of cooking on one muscle, versus that on cooking the retail cut, should be determined. A possibility is cooking the retail cut and analyzing one muscle only.

It is assumed that all laboratory controls possible will be used to obtain uniformity.

The quality of the food used will depend upon the objectives of the study. If the study is on the effect of storage on the quality of frozen or dehydrated food, one would probably start with high quality so that any decrease in quality would be measurable. In some types of studies one might well start with lower quality. For example, if one wished to determine the effect of freezing and freezer storage on the tenderness of meat, one would start with a less tender cut so that any tenderizing effect would be noticeable, which it would not be if the original meat were already tender. Perhaps in this instance one might need high quality for a reference standard.

Number of samples and replications

The number of samples that can be evaluated satisfactorily at a given time is dependent largely upon:

(1) Characteristics or factors being judged. In judging odor and texture, sensory fatigue does not set in so rapidly as it does in flavor judging. Therefore more samples can be judged at one time for odor and texture than for flavor. As many as 20 samples can be scored for color, 12 for texture, and perhaps only 4 or 5 for flavor.

(2) Intensity of flavor of food. If a food has a bland flavor, for example milk or bread, 5 to 10 samples may be judged satisfactorily. If the food has a strong flavor such as spice, fewer samples, perhaps 4 or 5.

(3) If the foods vary greatly in the characteristics being judged, more samples may be tested satisfactorily at one time than if greater discrimination is required.

(4) Number of characteristics to be scored. If 6 factors are to be scored perhaps 4 samples may be judged, but if only one factor is to be scored perhaps 24 samples may be judged. Research is needed to determine whether this is true.

The number of sessions which may be satisfactorily conducted in a given day is about four at the most, depending upon the nature of the test. Conditions which favor fatigue will reduce the number of samples that can be judged satisfactorily.

Three replications are necessary on a judgment basis, four are better statistically. Three replicates provide too small a number of degrees of freedom. If results are marginal, more replications are necessary. If the panel is small more replications may be necessary than if the panel is large. Research is needed to determine whether this is true. The number of replications necessary is determined somewhat by the purpose of the study, the time alloted before results must be reported, and by the money available.

Reference standards

The importance of reference standards has been mentioned several times. The use of more than one standard representing several levels of quality has been suggested. Reference standards in the medium level rather than the highest level have also been suggested so that the judge is not always scoring in one direction.

In long-term storage problems the difficulty of obtaining standard samples of changing biological materials was pointed out. It was suggested that a partial solution may be the use of one or more of the treatments as reference standards. This method does not, however, permit the panel to judge against an abstract standard but only provides a means of comparing the several treatments.

Wax models may be developed which can serve satisfactorily as reference standards in judging color, shape, and size.

Mildred Boggs reported that samples seem to undergo little, if any, change when they are stored in a well-insulated box with dry ice at $-70°$ C. The availability of good reference samples would do more than anything else to improve the accuracy of palatability testing.

Amount of information given panel

All of the information that you can give the panel members should be given so long as they cannot identify particular samples and do not become prejudiced. This is for the purpose of gaining the interest of the judge and to increase his ability to concentrate.

The judges should be told the reason for the study or what instigated it. They should be told the need for the information sought. They should be given the present status of information or a brief review of the literature. They should be told the plan of study, including its estimated length, and given information on the objective tests, particularly those that might be correlated with organoleptic tests for any characteristic. The amount of detail given would depend upon (1) the backgrounds of the judges, (2) the frequency of the tests, (3) the plan for continuing training and reporting to judges where their scores stand in relation to the group, and (4) the details of the administration of the tests.

Scheduling of samples

Milder flavors should be tasted first because stronger flavors are likely to affect the judging of subsequent samples. If this is always done, however, the judges may know it. Crocker suggests in his book, Flavor, that each judge arrange his samples in the order of increasing odor and then taste them in that order.

The judges do better scoring if the samples are not too disagreeable. Therefore if after 1 year's storage the samples are very disagreeable it is unwise to continue the storage study over a second year. You may lose your judges!

Flavor memory is more persistent than odor memory.

The time interval between samples should be long enough to permit the return of normal saliva flow, and also, as in the case of sulfite, to permit overcoming of the dulling effect of sulfite flavor on the judges' ability to detect differences. Research is needed on how long these time intervals need to be for different kinds and amounts of food.

Odor, color, and texture should probably be judged before flavor.

COMMITTEE: Grace Bennett, Ernest C. Crocker, Ivan D. Jones, Mary E. Kirkpatrick, Inez Prudent, Faith Fenton, Chairman.

Correlation of Sensory Tests with Chemical and Physical Tests

Discussion

BELLE LOWE: Several people have asked me how I've kept out of the argument so far and I have said that I was prepared to discuss one topic

and that most of the time I would stick to that. A topic like this cannot be covered adequately in a short period. As a beginning, I want to mention gadgets briefly. As a nation, we like gadgets — I have been guilty of inventing a few myself. We won't talk about the water pressure penetrometer used with jelly, and some of the other instruments. There are a great many gages, most of them not good. There are the colorimeters, electric currents used for various things, line spreads, height, volume, peroxides, free fatty acids, and so on.

I brought along some slides to show some correlations between physical tests and scores, between chemical tests and scores, and between histological tests and scores.

The first slide shows three objective tests and one subjective test for soft custard. The second column shows stiffness by ranking, the third is the MacMichael viscosimeter, and the fourth is the penetrometer. The first three tests were with fresh eggs 24 hours old, with three Haugh units, a high, medium, and low. The fourth test was on aged eggs.

The second slide shows the Haugh units with the viscosimeter readings for custards. This illustrates the fact that with stirring all the time, the viscosity goes down.

The third slide shows measurements of pH of egg white, using three fresh eggs with the different Haugh units. It illustrates the difference between pH of custard mix before and after cooking.

We measure tensile strength in angel cake, sponge cake, bread, rolls, and the like. The next slide shows an adaptation of an ordinary scale in graph form. The sample was put between clamps and sliced in a miter box to make it of uniform thickness, and of an hourglass shape. Now we see scores and tensile strengths on angel cake. This cake was mixed in a single batch, divided into four cakes, and baked at the same oven temperature but with different baking times. As you overbake egg mixtures, they tend to become tougher. The cakes baked a longer time showed correspondingly less tenderness on both the scores and the tensile strength readings.

Often our objective tests do not correlate with the scores of subjective tests given by the panel. It takes a lot of common sense in addition to statistical interpretations sometimes to find out the reasons. Very often I think the panel is more accurate. There are other tests where the objective test is more accurate. We have an example of that in the next slide, which shows results of some work we did this fall. We used right and left cuts from the same animal. One was supposed to be cooked to 90° C. and the other to 70° interior temperature, but both were cooked to 90°. The panel evaluated both the same on tenderness scores. The pressometer reading percentage, however, which measures the amount of juice that can be pressed out in 5 minutes under 250 pounds pressure, shows 28 percent for the right sample and 33 percent for the left sample. That is a wide enough difference in percentage so that we knew there must be some explanation for it.

We went back to the cooking and discovered that one sample was cooked 140 minutes and the other 92. In cooking proteins, when you reach the point where coagulation begins, there is an endothermic reaction and sometimes the food will absorb heat and stay at the same temperature for minutes, sometimes for hours, particularly if you are using a very low temperature. In this case, the pressometer was more accurate than the panel. Our panel has not been as good on the one

factor of juiciness as it has been on the other palatability factors we have scored. This is probably due to my not training them enough on it.

It is interesting to note that the pressometer test on the amount of juiciness agrees with the cooking loss. The meat cooked a longer time had a higher percentage of loss, which is to be expected. Since this loss is largely moisture, you would expect the pressometer test to agree there. The longer cooking time, in addition to volatilizing more of the liquid, may have caused binding of water for the right sample to a greater degree than for the left sample.

The next slide shows some of Inez Prudent's and Dorothy Harrison's work. The chemical work was done on two animals, the palatability and histological tests on four animals. The first slide is on animal 1. Column 1 is muscle, *p* stands for the shortest measure of the tenderloin, *l* for the longest; *l r* is the rib portion, *l l* is the loin portion, *s t* is the semitendinosus muscle from the round, and *s b* is the semimembranosus. *C* stands for the nitrogen collagen in percentage of the total nitrogen, *e*, the elastin nitrogen in percentage of total nitrogen, *t* is the tenderness score, shear, and the histology column is the last.

We have had the idea for a long time that when a piece of meat has a high collagen content, it is tougher. You will see that most of the way through, the scores and collagen pretty well agree.

The histology column is just a reading from the observation of the histological sample. When you use these very small samples, there is the likelihood of very big errors because of lack of uniformity throughout the meat.

There is one exception in agreement — the semitendinosus muscle with a high collagen content; the scores and the shear do not agree. We have had trouble with that muscle for 20 years. I think we are getting at it but we need to do a few more animals. The very high elastin content in that semitendinosus muscle, which is higher than any of the others, is reflected in the shear but not often in the score. We get similar results with poultry.

The next slide shows Inez Prudent's work. The muscles were aged different lengths of time — 1, 2, 5, 10, 20, and 30 days. The solid upper line shows average collagen content of the muscle in proportion of nitrogen and the lower line shows the elastin. The dots show the variation. Notice that there is no increase in collagen — it was not significant during 30 days of ripening. Animal husbandrymen have told me that the collagen content changes with ripening of meat, but this shows that we have to look for physical instead of chemical change. Notice that, of the different muscles, the semitendinosus is the one very high in elastin — the other four are not very high.

The next slide shows the first score card of animal 1, a good grade of steer beef animal, about 18 months of age. Animal 4 is an 8-year-old dairy cow. Again, we get about the same correlation, the semitendinosus again giving the "off" correlation between scores and shear and histology. Notice again that the collagen and elastin will go up for the older animal and not so much in the younger animal.

The next slide shows the graph. You will note that elastin content is less in the old dairy cow than in the 18-month animal, particularly for some of the muscles. The collagen content is much higher in the older animal.

The slide you're now seeing shows some tenderness scores and shear with poultry. The birds were aged for different periods of time and roasted because we could control the temperature easier with that method of

cooking. The birds in the first row were in the oven 10 minutes after killing, then 30 minutes, an hour, and so on up to 5 days. There were five birds in each group.

In judging tenderness, we counted the number of chews required to masticate a certain size sample. We took into account the amount of chewing surface of each person and the strength of the jaw muscles by establishing end points for changing the score. The end points vary with different kinds of samples. We find that with this method we get close correlation between the panels and the shear apparatus.

(Histological slides were shown and Miss Lowe stated that turbulence, waves, and longitudinal striation are all well correlated with toughness.)

There is a variation in the rapidity with which the changes come about in aging (and tenderizing) animals. Broilers tenderize more rapidly than fryers and roasters are slower than fryers. The old cow doesn't tenderize as rapidly as the roasters or the younger beef animal. There is also a variation from muscle to muscle. The white meat in poultry tenderizes more rapidly than the dark meat. There is also variation in individual animals. Another thing we discovered is that broiler white meat will tenderize nearly as much in 3 hours after it is killed as it is ever going to tenderize. The dark meat of a broiler will get fairly tender in 18 hours — enough to rate an 8 score on a 10-point scale. It takes the roasters about twice that time. We have found that aging poultry for 1 hour is equivalent to 1 day in beef. Although there are individual variations, with 5 to 10 birds the averages come out about the same each time.

A. KRAMER: The report I am going to make is based on work that was started at the National Canner's Association around 1940 and continued later at the University of Maryland. The purpose was to replace the panel with objective tests. We hoped through objective tests to improve on precision, since results can be repeated better. We were using groups of canners as judges. They were accustomed to evaluating quality in terms of standard, extra standard, and substandard so we set up scores from 1 to 4. They evaluated a fancy sample at 4, extra standard at 3, standard at 2, and substandard at 1. After a while, we found that they were using plus and minus with these scores, in an effort to evaluate with greater accuracy. We therefore expanded our scale so that a high fancy was 10, medium fancy was 9, low fancy was 8, extra standard 7, 6, 5, standard 4, 3, 2, and substandard was 1.

The size of the panel made up of canners ran anywhere from 15 to 50 persons. They were supposed to know the product; the commercial scores and the results were supposed to be suitable for immediate use in the trade. We discovered that the people who were producing the food product were not always the most precise judges, so we set up another type of panel. This panel consisted of 8 to 12 experts including usually 2 people from the Inspection Division of the Production and Marketing Administration, 2 from the Horticulture Department, 2 from Home Economics, 2 from Agricultural Marketing, and 2 others.

We found that with a series of samples covering the entire commercial range, correlations between the consumer panels would average about 0.8.

QUESTION: May I ask the size of the consumer panel?

A. KRAMER: These panels of canners ranged from 15 to 50. With the "experts" panels, correlations ran as high as 0.9 or even better. As a result of this experience, we have used arbitrary methods to determine when an objective method was closely enough correlated with a panel

result so that it could replace the panel. We concluded that 0.8 correlation was sufficiently high. We assumed that if 2 panels would not agree any better than that, a series of physical or chemical values should not be expected to be much higher in correlation. We eliminated results that correlated below 0.7 and considered results with a correlation between 0.7 and 0.8 as worthy of further consideration.

In objective tests, it is important to define as exactly as possible, and to isolate as far as possible, the particular factor that is to be measured. Over-all quality tests are never accurate. Often the objective method is too long and too tedious and the various factors that have to be judged require hours and hours of chemical and physical procedures to arrive at the results for one sample. In such cases, the panel is obviously a faster and cheaper method of evaluation. Objective tests are helpful in setting up reference samples.

Another use for objective methods is in case of borderline samples. We can run the sample on the machine and see what the machine says. Objective methods, too, can be used to eliminate or hold constant some factors while other factors are being measured. For example, we may wish to evaluate different varieties of corn of the same maturity. If we use moisture content as a measure of maturity, we may decide to take 70 percent moisture as the point at which we will get all of the varieties to be tested. Or, if we can't obtain our samples at the same stage of maturity, we can determine the moisture content and perhaps by some covariance procedure, eliminate the maturity effects in measuring flavor differences.

Now, as to our actual measurements, I have divided them into three general categories. The first is appearance. Appearance factors such as size, shape, and pattern, all can be measured satisfactorily and rather easily with objective tests. Light measurements can be divided into two factors, one being gloss and the other color. Gloss can be measured with the goniophotometer. Measurements of gloss are important with such products as tomatoes and raw apples.

Color can be measured most simply by referring to a color dictionary. That is frequently done but with poor results. Somehow, the color dictionary does not give us just what we want. We have two general methods for measuring color. One method is with reflectance measurements, where we can use a disc colorimeter, usually with the Munsell system. With this instrument we can combine color cards in the form of discs which are spun around rapidly and get the color which matches that of the food product. Values are obtained in terms of hue, chroma, and value. Hue is the actual dominant wave length, that is, whether the color is red, yellow, or green; chroma refers to the intensity of the color; and value denotes the amount of darkness in the color complex. The cards have to be compared visually, so that perhaps more objective results can be obtained with the General Electric spectrophotometer with a tristimulus integrator attachment.

There is one drawback to this instrument. The General Electric spectrophotometer costs about $8,000 and the tristimulus integrator costs about $6,000. There is another instrument, however, which I believe you will be interested in knowing about if you haven't already heard about it — that is the Hunter color difference meter. We are working on it now. This instrument replaces an assortment of color cards with photoelectric cells. There are five photoelectric cells with proper filters in front of them. The instrument provides three different measurements with the three different attributes of color on a three-plane

dimension. If we are prepared to spend the money for the instrumentation and do a certain amount of preliminary work in preparation of the sample, we can get objective measurements of color.

There has been a lot of work done on instruments to measure the next group of factors, which I will call kinesthetic, a word borrowed from Mr. Crocker. As Miss Lowe said, we have a tremendous number of gadgets for various purposes, and like Miss Lowe, I can take part of the blame for producing some of them. We are now reaching the point where we have so many texturometers, tenderometers, fiberometers, pressure testers, fiber pressure testers, etc. that we'll have to move out of the laboratory soon to make room for all of the gadgets if we are working on more than one or two products.

We are, however, working on an instrument that should be useful for measuring a lot of products and a lot of factors. Fundamentally, we are not measuring too many different things. We're measuring either shearing force or we're measuring pressure and we can also add a few other things, perhaps penetration, and cutting force, which is an attempt to reproduce the effect of the fork. For example, when you eat asparagus, you want the fork to cut through it, otherwise it is too fibrous. Our idea is that we want to measure just a few things and there is no reason why we shouldn't do our measuring with one instrument which supplies the force and the measuring scale. A series of sample boxes is used for the different kinds of measuring. We think the instrument we are working on will do these things. Our Engineering Department has drawn it up and is making it now.

We plan a sample box for measuring tenderness. Peas, for example, will be measured for their resistance to shearing with a series of bars one-eighth inch thick. If we wish to measure the firmness of a marachino cherry, we would place it on a plate where the force required for this plate to come within one-eighth inch of the platform would be measured. We will probably have a series of gages — one gage may go up to 1,000 pounds, another to 50 pounds.

This machine can be used also for measuring succulence. We will probably hinge a sample box so that it will be on a horizontal basis and the juice will drip from the sample.

MILDRED BOGGS: May I ask how large a sample the machine measures — one pea or 200 grams?

A. KRAMER: The sample box, as I remember, is about a 2½-inch cube. That reminds me that many of the instruments are designed to measure only one unit, and from a practical standpoint that is not very useful. It can take all day to measure a representative sample if we measure one cotyledon of one pea at a time. Another disadvantage of some of the instruments we now use is the cost and the difficulty of comparing results from different machines. For example, the tenderometer we use for measuring tenderness of raw peas weighs about 600 pounds. It costs about $800 and its only purpose is to measure the tenderness of raw peas. It gives values in terms of shearing force in pounds per square inch but you can't be sure how the results compare with results from another machine.

This brings us to the third general group of factors — flavor measurements. There, objective measurements have little to contribute. I noticed that Miss Lowe also was very quiet about this group of factors. Of course it is easy to measure salt and sugar content of a product, but very difficult to get an over-all impression of flavor. Some work is being done at Ithaca on measuring the essential flavor of peas. The instrument

used by the Bureau of Fisheries, called the stinkometer, which measures the results of putrefaction may be of some value. We have done some work with a rapid method of determining the free amino acids. We find that the longer peas are held, the higher become the values, and also the younger the peas, the higher the values, so one just about cancels the other. It is difficult to know how to interpret the results.

I am going to close by saying that in correlating many of the panel results with objective results, we get a logarithmic curve. For example, if we correlate the viscosimeter (Stormer) values in terms of seconds per 100 revolutions and panel values for consistency of cream-style corn, we come out with a curve that is surprisingly logarithmic in nature. If we measure the relationship between fiber content and fibrousness of asparagus as determined by the panel, we also get a practically perfect logarithmic curve. We are now automatically converting one set of data into logs and getting the correlation that way. It is surprising how good the correlations are if we handle them in this manner.

MILDRED BOGGS: We have perhaps five people working on chemical and physical tests to one on palatability by panel methods, so practically all of our panel samples are measured by a number of objective tests. I don't know of any objective tests that will measure quality under all conditions of treatment. They apply under a certain set of conditions and not under another set. We find them extremely useful, we couldn't do without them, but it usually takes a long time to develop an objective test and determine its limitations.

I think we have to watch our correlations. You can get a beautiful correlation by simply throwing in a bad enough sample and a good enough sample. You may have a lot of dots widely scattered in the 5–6 area and you throw in a few scores of 1 and 10 and the test is worthless in separating score 5 material from 6 or, more specifically, it is useless for Federal grades.

BERNADINE H. MEYER: I have a very simple point to make. The exhibit (exhibit 15) shows the relationship of flavor scores on one set of sponge cakes to the peroxide number. The panel was small — seven people of whom four had previous experience on two sponge cake panels while three had no previous experience on a sponge cake panel. We could not offer during the training any samples with the particular off-flavor expected to develop during storage of the sponge cake, so the training had to be given to the new people with the normal flavor of fresh sponge cake.

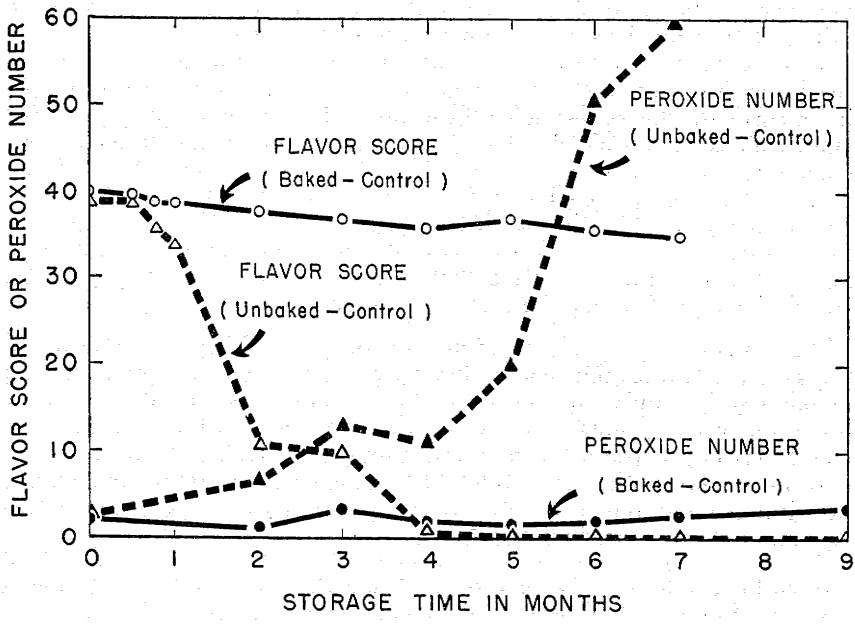
As you see from the exhibit, there was good agreement between the flavor scores and peroxide values. There is a marked development of off-flavor indicated by the scores, along with a definite increase in peroxide values. Our panel, even with limited experience, detected the off-flavor and judged the cakes inedible before the increase in peroxide was great enough to have significance. This test points out the agreement but emphasizes the importance of the panel in flavor evaluation.

MILDRED BOGGS: We use the Hunter reflectance instrument a great deal and find it very promising. Dr. Eastmond, who does our color work, prefers it slightly to the Hardy because it takes a larger sample.

CLAUDE H. HILLS: It seems that what we are trying to determine in tenderness scoring is how much energy it takes to chew up the food we eat. I don't think the shear measurements give us the right answer. I don't think the penetrometer gives the right answer. I wonder why

RELATION OF PEROXIDE NUMBER TO FLAVOR SCORE OF SPONGE CAKES

University of Tennessee College of Home Economics



someone hasn't developed an instrument that depends on grinding with dull blades.

ERNEST C. CROCKER: Dr. Collins of General Foods Laboratory has the answer for what you want. He has developed a complete set of false teeth that do the chewing. He can get a numerical answer for any sample tested that is far more reproducible than results of any panel member.

Committee Report

The advantages of using objective tests are evident. A physical or chemical method may be superior to an organoleptic method in precision but not in accuracy. An ideal method is extremely rapid and accurate. One method to reflect over-all quality is practically impossible, but the factor which is most important or the "worst offender" may be measured and used as an indicator of over-all quality. Generally, the chance for success of an objective method improves with the ability to isolate and define clearly each factor of quality.

Methods which are very highly correlated with organoleptic evaluation, and are not time consuming, may be used to replace organoleptic evaluations. Other methods which are too time consuming may be used for reference samples, for panel instructions, to settle borderline disputes, or to eliminate the effect of other factors occurring simultaneously.

Objective methods are available for measuring such appearance factors as size, shape, and pattern. Gloss-measuring instruments, such as goniophotometers and glossmeters are available but practically no use has

been made of them in testing food products. Color measurements may be made by the use of disc colorimeters and recorded as Munsell notations. Many detailed procedures are available. Transmission methods involving the extraction of characteristic pigments with proper solvents and measurement of transmittance of the classified pigment extracts are available for a large number of products. The use of the Hunter color-difference meter is under investigation and is very promising.

A large number of gadgets have been developed for measuring specific characteristics such as shear or tear strength, viscosity, pressure, and penetration, for a large number of food products. It was suggested that a more basic approach is needed whereby a universal multipurpose unit might be developed. This instrument would supply the power, measuring units, and a series of different types of sample boxes which may be fitted into the instrument for specific purposes such as measuring shearing force, pressure, penetration, or juiciness.

In connection with objective tests, chemical tests such as those for collagen and elastin nitrogen, and peroxide numbers and histological studies were mentioned.

Objective methods for determining flavor are limited to simple systems, where a single or perhaps several well-known constituents are responsible for the flavor or off-flavor. For many foods where flavor is the result of many interacting components, the identity of many of which may not be known, an objective determination is not now available.

Regression curves for many sets of data involving objective and organoleptic measurements are logarithmic in nature. A correlation coefficient of 0.8 or higher between an objective and organoleptic test is required before the objective method is considered to be a satisfactory replacement for the organoleptic, provided the set of data covers thoroughly the commercial range and does not consist of an excessive number of extreme samples.

In summary the committee emphasized that in comparing subjective and objective tests it is important (1) that the objective method and the scoring panel measure the same palatability characteristics, (2) that there is accuracy in recording data and in the use of the machine, and (3) that the design of the experiment, both as to sampling methods and procedures for objective and subjective tests, is carefully planned, as the validity of the results is to a great degree dependent upon these factors. The planning should include the statistical design, obtaining the samples, holding or storage method, temperature of testing, etc.

Realizing that there is dire need for the standardization of methods, the committee recommended that directions for objective methods which have been worked out at various laboratories be assembled and published in mimeographed form to be used as a reference tool in laboratory procedure. Recommendations for the use of each test, its limitations, and the meaning of the results should be included.

Use of objective tests worked out in one laboratory should be used in additional laboratories to evaluate the validity of the test and perhaps a referee system developed therefrom. Many data now available should receive additional study in an attempt to show the relationship between objective and subjective tests.

COMMITTEE: Amihud Kramer, Beatrice Mountjoy, Helen J. Purinton, Belle Lowe, Colburn C. Fifield, Gladys Vail, Chairman.

Discussion

GERTRUDE COX: Whether you are evaluating effects on palatability of foods, of storage, preparation of food, cooking time, varieties, maturity, dilutions, insecticides, or containers in which the food is kept, your results vary. This variation introduces a degree of uncertainty into any conclusions that are drawn from the results. Even after having several judges well trained with the power to discriminate, score, or rank material, you do not know how much the results would be changed if you had another group of judges. If you had a perfect measuring device for evaluating palatability of food, most of your problems would be easy. Obviously, you cannot expect to secure true values, but limits can be found with the probability of enclosing the true difference or true value with a specified degree of certainty.

The statistical solution to the problem of estimation consists of a statement that the true mean lies between certain limits, plus a probability that the statement is correct. Research workers have found that this type of information does provide a basis for action.

Let me repeat, variability in results is typical in experimentation. Because of this, the problem of drawing conclusions from the results is a problem of induction from the sample to the population. The statistical theory of estimation provides solutions to this problem in the form of definite statements that have a known and controllable probability of being correct.

Now let us discuss the initial steps in planning the experiment. Since the inferences that can be made depend upon the way the experiment was carried out, statisticians should have a detailed description of the experiment and its objectives. Our participation in the initial stage of experiments in different areas of research leads to a strong conviction that too little time and effort are put into planning. You get the money and must start right away and you expect to plan the experiment as you go along. This does not work. The statistician may contribute to the planning simply by getting the investigator to explain clearly why he is doing the experiment, to justify the experimental treatments whose effects he proposes to compare, and to defend his claims that the experiment when completed will enable his objectives to be realized.

It is a good practice to make a written draft of the proposal for any experiment. This draft will, in general, have three parts: A statement of the objective; a description of the experiment, telling about the experimental treatments, number of judges, and method of scoring that will be employed for this particular test or experiment; and ending up with an outline of a proposed method of analysis of the results.

The most common faults of projects are their vagueness and number of objectives. To avoid having too many objectives and trying to do too many things in one experiment, it might be advisable to have major and minor objectives.

When you have stated the objective, selected the variables to be measured, the treatments to be used, pointed out the laboratory difficulties, specified how much work you can do in a day, and what kind of instruments you are going to use, then the statisticians can give you suggestions regarding a design or plan for the experiment. We should be able to tell you which is the most efficient method to carry out that experiment. We might say, "Try this particular design and we know

you will get results that will give you an unbiased estimate of error and effects. We are not sure this is the most efficient method but until we accumulate experience either by designing experiments to give us the information, or by getting together results from a great many people, we cannot tell you which is the most efficient design."

I spoke of variation, that common element that we have in our experimental work, which we call experimental error. Now, I want you to think about your experimental unit. An experimental unit is a group of material or an individual object to which a treatment is applied in a single trial of the experiment. We must agree that certain sources of variation are important and we must know what material you are going to use before we can determine the experimental unit — whether one experimental unit is the score of a judge, or whether it is the average score of a group of judges, or whether one judge may be an experimental unit each time he scores the material.

We have spent a good deal of time the last several years seeking methods to increase the precision of experimental error. Increasing the precision does not always mean making a smaller experimental error. Sometimes you are using too small an error for your experimental situation. To increase the precision of an experiment, we can increase the size of the experiment by adding more judges, doing it more days, or sometimes by adding more test material. Using more treatments within this one experiment may give you more degrees of freedom for a better estimate of the experimental error. Or you might increase the precision of your experiment by refining your technique. That is, of course, what you have been discussing this week. Another way of increasing the efficiency or precision of the experiment is by the selection of your experimental material, that is, selecting your judges or the material that is to be judged. A good deal needs to be known about the product being brought into the laboratory for the taste panel.

Often you can make objective tests as well as your subjective flavor test and in the analysis of covariance make an adjustment for the variation in this objective factor which is related to the variable being studied. Any time that you can take measurements on related factors, even though you may not think they are too good, if they have a relationship to the variable you really want to measure, you can often raise the efficiency of the experiment 30 to 40 percent with little extra cost.

If a control is required, it should be an integral part of the experiment so that results from the control are directly comparable with those of the other treatments. (I think we can do more thinking about what is your control or standard. We might seriously consider having the control at 80 percent or even 50 percent instead of 100 percent.)

I want to discuss a few experimental plans. Those plans are listed in the very general article that I gave at one of your Chicago meetings and is published in the Journal of Home Economics (Nov. 1944, p. 575). The book on experimental design that Professor Cochran and I are completing will have a rather complete record of the designs that we have found useful. Mathematical statisticians have prepared a whole series of designs that are not yet published. We will try to evaluate their usefulness before we suggest them for general use.

Planning an experiment consists of putting restrictions on the way treatments are assigned to the experimental units. We have first the completely randomized design, in which there are no restrictions except that you do it completely at random. I find this design almost impossible to use in experimental work, especially where you are going to super-

impose a restriction by the way you handle the material. You may use it in the chemistry or physics laboratory where exceedingly homogeneous material is being used, and many tests can be made in a short period of time. The most commonly used is the randomized block design. If an easy design will do, never use a complex experimental plan. The randomized block design can become complicated. The original design was straightforward in a randomized arrangement in the field but, as you began to work with the material, you introduced restrictions. The experiment must be analyzed according to the way the research person did the job and it is very important to know how he handled each step.

Consider an experiment to evaluate three packaging methods, A, B, and C. You are using five judges and you want them well trained if you are going to use them as machines to do the measuring.

		Method of Packaging		
Day and Judge		A	B	C
Monday	1	—	—	—
	2	—	—	—
	3	—	—	—
	4	—	—	—
	5	—	—	—
Wednesday	1	—	—	—
	2	—	—	—
	3	—	—	—
	4	—	—	—
	5	—	—	—
Friday	1	—	—	—
	2	—	—	—
	3	—	—	—
	4	—	—	—
	5	—	—	—

The material was packaged at different times. On Monday each of the five judges was handed three materials at random. I am assuming that there is field variation so that the material given to the judges on Wednesday differed from that received Monday. The analysis of variance follows:

Source	*Degrees of freedom*
Days	2
Methods	2
D × M (error)	4
Total	8

You will note that the experimental unit is the average score for the five judges. If your judge-to-judge variation is small, you do not need many judges at a time. You need to repeat your experiment more times in order to get a better estimate of the variation in the material being scored. Also we need more than 4 d.f. for error. So instead of five judges on 3 days it might be more efficient to use three judges on 5 days.

Now, I am going to make a slightly different assumption. I have two sources of random variation, (1) judge to judge and (2) material to material.

I am assuming that I have a representative sample of this material by taking three samples. I also want to make the assumption that I have a representative sample of judges. These judges reasonably represent a population of consumers about whom you are concerned. I want to be sure that the two kinds of variation (material and judge) get into my error term because I know they are in my means. The analysis is:

Source	Degrees of freedom
Days	2
Judges	4
D × J	8
Methods	2
Experimental error	28
Total	44

The experimental unit is a single score of a judge. In the previous analysis, the mean score of five judges was the experimental unit.

Now let us see if we cannot secure more information from the first experiment. I have a judge difference. The failure of judges to perform the same on different days may be valuable technique information. Judge × method interaction certainly is of interest, to see if judges are scoring various methods consistently. Likewise, judge × day interaction would be of interest. This analysis takes the form of:

Source	Degrees of freedom
Days	2
Methods	2
D × M	4
Judges	4
J × D	8
J × M	8
J × D × M	16
Total	44

The Latin square design is used when it is desirable to place two restrictions on the assignment of the treatments to the experimental units. The order of presenting the material to the subjects is one restriction. This means that only one sample is given to the judge at a time.

	Order of Testing				
Judge	I	II	III	IV	V
(1)	*A*	*B*	*C*	*D*	*D*
(2)	*B*	*D*	*A*	*C* Rep. I	*C*
(3)	*D*	*C*	*B*	*A*	*A*
(4)	*C*	*A*	*D*	*B*	*B*
(5)	*D*	*B*	*A*	*C*	*C*
(6)	*B*	*C*	*D*	*A* Rep. II	*A*
(7)	*A*	*D*	*C*	*B*	*B*
(8)	*C*	*A*	*B*	*D*	*D*

In the first replication, there are four judges and four treatments (*A*, *B*, *C*, and *D*) to be assigned to the judges in a prearranged random order. You will note that all judges have had all four treatments. Also every treatment has appeared first to one judge and every treatment has been the last to be tasted by some judge. In replication II, the same judges may be used again, or judges 5, 6, 7, and 8 may be used as indicated in this example. The analysis would be:

Source	Degrees of freedom
Judges	7
Order	3
Treatments	3
Error	18
Total	31

In order to measure the carry-over effect, another column, V, is added: Each judge is given a duplicate sample of the material assigned to him in

column IV. Now every treatment is preceded by itself and by every other treatment in each replication. The analysis of this design is too involved to be presented here.

I would like to present one incomplete block design but I am not advising its use unless you have access to a statistician, or are quite familiar with the characteristics and analysis of this design.

	$t = 9$	$k = 4$	$r = 8$	$b = 18$
Judge	(1) 1234	(7) 1278	(13) 1468	
	(2) 1489	(8) 2389	(14) 2679	
	(3) 2568	(9) 4679	(15) 3678	
	(4) 1256	(10) 1357	(16) 1369	
	(5) 1579	(11) 2459	(17) 2347	
	(6) 3589	(12) 3456	(18) 4578	

t = treatments, k = number of units given to judge at one time, r = replications, b = number of judges.

This requires 18 judges or 6 judges used on three different days. We now have every pair of treatments scored by three judges, or every treatment appears with every other treatment in a group of four, three times. Individual differences are removed in the analysis when comparing treatment means. We have nine treatments but are asking a judge to score only four at a time instead of nine.

Here are three designs — randomized block, treatment paired with control, and incomplete block type of experiment.

I. Randomized block

C	*A*	*B*	*D*
B	*A*	*C*	*D*
D	*C*	*A*	*B*

II. Paired with control

A	*B*	*A*	*B*
A	*C*	*A*	*C*
A	*D*	*A*	*D*

III. Incomplete block

A	*B*	*B*	*C*
A	*C*	*B*	*D*
A	*D*	*C*	*D*

To get a line on the comparative value of the three designs, I would suggest giving each judge a different design as follows:

Judge				
	(1)	I	II	III
	(2)	II	III	I
	(3)	III	I	II

Take two more Latin squares changing the order of arrangement:

(4)	II	I	III	(7)	III	I	II
(5)	I	III	II	(8)	II	III	I
(6)	III	II	I	(9)	I	II	III

You have information on your judges, definite information on your methods, and an error term.

In conclusion — it seems that you need to learn more about the sources and size of experimental variation. Experiments could be conducted to test your judges, your methods, and the designs. We need to know how many judges are needed, the relative value of different methods of evaluating results, and the efficiency of various types of designs.

Committee Report

Variability in results is typical in experimentation. Because of this, the problem of drawing of conclusions from the results is a problem of

induction from the sample to the population. The statistical theory of estimation provides solutions to this problem in the form of definite statements that have a known and controllable probability of being correct.

Statisticians are often asked for advice in making inferences from the results of experiments. Since the inferences that can be made depend on the way in which the experiment was carried out, the statistician should have a detailed description of the experiment and its objectives.

Accuracy refers to closeness to "true" measure; to the representativeness and absence of bias in the experiment.

Precision refers to repeatability of measurement as measured by the expected range of variability in results of a series of similar experiments.

Most of the discussion regarding devices, replication, additional measurements, and skillful groupings relates to the precision (repeatability) of the experiment.

Treatment paired with control represents inefficient use of experiment and worker. To test the relative effectiveness of the designs, namely, (1) complete randomization within a block, (2) randomization by pairs which include the control, and (3) incomplete block, it was recommended that different methods be assigned to each judge on different days, using three Latin squares. For example, three different scoring sheets could be used at the same time. For Method I, use 1 to 10 scoring scale; Method II, 0 to +5 and −5; Method III, 0 to 5.

Recommendations

Get into the material and come up with some answers to the questions that have been discussed in these sessions.

Learn more about sources and size of experimental error, and number of judges needed.

Specify the size of the true difference which the experiment is to detect by means of a test of significance, or specify how closely the true difference is to be estimated, by stating the width of the confidence interval.

COMMITTEE: E. J. Koch, D. D. Mason, Isabel Noble, Andrea Overman, Gertrude Cox, Chairman.

Methods of Analyzing Data

Discussion

MARY L. GREENWOOD: Miss Cox has pointed out that the analysis to be done should be planned along with the experiment, rather than decided upon after the experiment is finished. The type of analysis is somewhat determined by the way the experiment is set up. A plan for analysis is needed in the beginning, even though it may be altered during the progress of the experiment.

I meant to preface my remarks by saying that what I am giving represents the opinion of the committees on experimental design and analyzing data, so it includes the thinking both of the statisticians and those of us who are in tasting work.

Sometimes the results of an experiment are so obvious that perhaps analysis is unnecessary. I will discuss analysis of variance which has wide use. Through this method of analysis, we can determine differences in means and also get an estimation of variance. If we want to

know how well our tasters can taste, we can discover how much an individual deviates from the panel mean and also whether he can repeat his own judgment. We can get a lot of information from interactions, which was pointed out by Miss Cox this morning. In experimental work with methods of treatment, methods of cooking, or other methods, we are interested in seeing whether the reactions are the same in every case or whether we get certain interactions. It has been suggested that if we don't get the expected answer, we have used an incorrect analysis. We need, then, to go back and think through what it is we are looking for, and to determine whether we have properly interpreted the data.

I would like to say a little about ranking. We used the randomized block that Miss Cox showed us this morning in our work on blueberries. You can get a tremendous lot of information this way. We used the incomplete block, where we set each one off against every other sample in groups of either four or five. We are able to get 15 judges who carry through the panel. We transformed our ranks to scores, although we are told that doing so makes a nondiscreet variable out of what is really a discreet variable. There may be a little error there although the statisticians say it is not too serious a matter. They tell us we can use these ranks when we don't know the distributions or when the distributions might be of a nonnormal nature, where we might get only one side of the curve although we might be getting these extremes. In such cases, the ranks are better used than the scores. Dr. Mason has agreed to discuss the least standard difference.

D. D. MASON: In analysis of variance, one of the first and most useful items of information we get is the test of significance of difference in treatment means. We are interested in determining just what means do differ at a probability level from other means. For example, where we have only two treatment means, A and B, and make the F test, we have only one possible comparison to make. If we get a test of significance at the 5- or the 1-percent level (or whatever we might be using), we're fairly sure of our results. Even here, of course, we might want to calculate our standard error of the mean and standard error of the differences in order to determine with what precision we did determine these differences.

A common mechanism used when we have more than two means — let us say A, B, and C treatment means — is to calculate what we call the least significant difference or as commonly abbreviated, L. S. D., which is our standard error of the difference times our t at the probability level that we are using with the appropriate degrees of freedom . This, particularly in our plant science work, has come into fairly common use and sometimes erroneous use, arising from the practice of setting up the treatment comparisons after we have inspected the data. For these F and t tests to be valid, strictly speaking, we should plan the comparisons before we inspect the results.

After we inspect the results, we will note that some treatment differences are larger than others. For example, we may have three values. A might be a great deal larger than C after we look at the results, possibly because of two factors. One factor might be that there is actually a large real effect between A and C and also there may be a fortuitous combination of errors that would make this apparently larger than it actually is. Incidentally, this point is covered in Cochran and Cox's textbook (Experimental Designs. 1950). Quoting from some of their probabilities, let us say we have three treatment means and we calculate an L. S. D. using 5-percent point. If we use this L. S. D. indiscriminately to compare

the means of A, B, and C, we are not any longer testing at the 5-percent point, that is, if we compare the highest with the lowest in six means. It is raised in that case to about 13 percent. With 6 treatments and comparing the highest mean with the lowest mean, the figure is about 40 percent; with 10 treatments, 60 percent; and with 20 treatments, 90 percent. In other words, if we have a list of 20 means and compare the highest with the lowest, it is fairly probable that we will get a difference that is greater than the calculated least significant difference, even though these 20 means may be from the same population.

The L. S. D. is frequently used to evaluate differences between treatment effects. It is like a measuring stick that can be placed between various treatment means for comparison, but we should be aware of the limitations of its use. As already mentioned, we can have a control or standard or check, whatever we may want to call it, and the use of the L. S. D. as calculated is presumably valid for comparing any test treatment against the control, because this is a part of our planned comparison. As was indicated by Miss Cox this morning, it is better to have our control near the center of the scale rather than at either end, possibly because of the disturbance of the distribution and also the fact that we won't be measuring all in one direction or the other. I want to point out that while we can't ignore unexpected results or unexpected differences that come up, we should consider them an indication that further investigation is needed rather than accepting them as the final result.

MARY L. GREENWOOD: Analysis of covariance has been mentioned here indirectly but I think the closest we can get to it this morning is in the discussion of correlation and regression. Lyle Calvin has agreed to talk to us about that.

LYLE CALVIN: This has to do with regression in connection with some data which Dr. Kramer has obtained. We've done a little work to see what the weight would be of the different factors that he has used in green beans. He measures color, maturity, fibrousness, and flavor on a 1–10 score. Independently and before any of these scores were taken, he obtained an over-all grade from his panel with a range of 1 to 4. In order to know what relationship each of these variables — color, maturity, fiber, and flavor — has to the over-all grade, a weighting system is needed to indicate the importance of each of these factors in predicting the over-all grade. The method commonly used in obtaining such information is regression. Since we had more than one variable, we used multiple regression and got a percentage weighting as follows: 5 for color, 46 for maturity, 17 for fibrousness, and 32 for flavor. These weightings were similar to those assigned by the experimenters independently at the beginning of the study.

GERTRUDE COX: I want to say something here about correlation. Someone reported yesterday that he was getting a correlation of 0.8 evaluating judges. That means that 64 percent of your objective score is common to your subjective flavor score. I think that is a more realistic way to say it. The square of the correlation gives you the amount of information that is common to the two variables. Your statistics may give you a correlation of 0.4 which is highly significant, but for predictive purposes, even though the statistics say it's highly significant, you would not use 16 percent common information on an objective test with a flavor test, for predicting. Even though the statistics give significance, you can't go just by your statistics. You've got to use common sense in inter-

pretation to see if the relationship is high enough to make it worth while to use the objective score instead of your flavor score.

LYLE CALVIN: In that connection, we might point out that there are really two types of significance. One is the statistical significance and the other is the biological or economic importance or significance of the data. You may have a difference at a 5-percent level that may not mean a thing. It may be one-tenth of what you're measuring whereas you'd have to have a much larger difference before it would mean anything. You may be able to measure it, and yet it may not mean anything, economically.

MILDRED BOGGS: Will you tell us what interaction is and why?

LYLE CALVIN: I'll illustrate with judges and methods. Take these values for A, B, and C samples for one judge and these values over here for the other judges (illustrating on board). The differences between A and B might be the same for judge 2 as they are for judge 1. If those differences are constant, the measured interaction will be zero. Interaction is the failure of one judge to be the same from method to method as the other judge.

MILDRED BOGGS: What causes this?

GERTRUDE COX: The failure of the differences in scores of the judges to be the same and also the failure of the differences between pairs of samples to be the same produces the interaction. For example, Mary Greenwood liked this candy better one way and I liked it the other. If we put our scores together, we'd have an interaction.

Committee Report

The analysis of an experiment should be considered when the experiment is planned. If the answer to an experiment is obvious, an elaborate analysis is unnecessary. The techniques that were discussed in this section were those that appeared to be in use by many of the participants and warranted further elaboration.

Analysis of variance

Analysis of variance is highly informative. It gives valuable estimates of the variances of factors under study as well as tests of significance. It is useful in selecting judges on the basis of erratic or consistent behavior which is a better basis of selection than mere disagreement with the panel. Interactions permit obtaining such information as the significance of treatment values within different storage periods. This information is lost if only paired means are studied. If the expected answer is not forthcoming, the analysis of variance technique may not have been applied correctly.

Ranking

Ranks can be analyzed by analysis of variance after they have been transformed to scores in terms of the normal deviate, as advocated by Fisher and Yates (Statistical Tables, 1949). The use of ranks lessens the influence of extreme values. Ranking is used when the distribution by using scores is (1) nonnormal or (2) not known. Incomplete block design was used in a study on blueberries.

Least significant difference

The L. S. D. is a measure by which the significance of the difference between treatment means (say A, B, C) can be ascertained. It answers such questions as "Is the difference between A and C real or fortuitous?" Since the L. S. D. is an average difference necessary for significance over all possible pairs of means, the use of it requires careful judgment. For pairs of means from near the opposite extremes of an array, the true L. S. D. is greater than the calculated L. S. D. For pairs close together it is smaller. For testing differences between the greatest and smallest means by use of the calculated L. S. D. (at 5-percent level), the actual probability level increases rapidly with the number of means in the array:

Number of means	*Actual probability level*
2	5 percent
3	11 percent
6	39 percent
20	90 percent

Correlation and regression

Green beans were graded on a scale 1–10 for color, maturity, fibrousness, and flavor. The samples were also given an over-all grade on a scale 1–4. The relationship between factor and over-all scores was studied. Weights on a 100 percent scale were determined by multiple regression analysis for each of the four factors considered as components of the over-all score. It is interesting that the computed weights were similar to those assigned by the experimenters independently.

Regression analysis is preferred to correlation analysis. The regression coefficient, by determining the slope of a line, makes it possible to estimate the degree to which one variable affects another. However, the square of the correlation coefficient has the following useful property: If r equals 0.8, then 64 percent (8 squared) of the variation of y is explained by x. An r of 0.4 (16 percent) might be significant in correlation but not useful in prediction.

COMMITTEE: Lyle Calvin, Elsie F. Dochterman, Mary L. Greenwood, Chairman.

LITERATURE CITED

(1) ANONYMOUS.
1937. FLAVOR IN FOODS. (Symposium.) Food Indus. 9: 314–315, 346.

(2) _____
1941. SENSITIVENESS OF TASTE JUDGES. Flavours 4 (4): 18–20, illus.

(3) _____
1943. METHODS EMPLOYED IN THE LABORATORY ANALYSIS OF DRIED AND FROZEN EGGS. 4 pp. (U. S. Dept. Agr., Food Distribution Admin.) [Processed.]

(4) _____
1943. DRIED WHOLE EGG. INFORMATION RELATING TO THE QUALITY AND SHELF LIFE WITH A LIST OF SUGGESTED RESEARCH PROJECTS. Coordinated Dried Egg Res. Program Rpt. 1, [18 pp.], illus. (Rev. ed.) (Natl. Com. on Poultry Prod. Res., Ames, Iowa.) [Processed.]

(5) _____
1948. HOW DO WE MAINTAIN SPONSORED BRAND QUALITY? HERE'S THE PART PERFORMED BY HALLMARK TESTING SERVICE. Safeway News 3 (8 and 9): 12–15, illus.

(6) _____
[1949.] FLAVOUR ACCEPTANCE. Flavours, Fruit Juices and Spices Rev. 11 (3): 9–16. (Reprinted in part from Amer. Perfumer and Essential Oil Rev. 52: 133, 135. 1948.)

(7) _____
[1949.] TASTE TAKES A TRIP TO THE LABORATORY. Laboratory 19: 27–31, illus.

(9) ALEXANDER, L. M., CLARK, N. G., and HOWE, P. E.
1933. METHODS OF COOKING AND TESTING MEAT FOR PALATABILITY. Sup. to Natl. Proj. Coop. Meat Invest., 36 pp., illus. (Rev. ed.) (U. S. Bur. Home Econ. and Bur. Animal Indus.) [Processed.]

(10) ANDERSON, E. O., DOWD, L. R., and STUEWER, C. A.
1937. RELATION OF ACIDITY OF MILK TO OXIDIZED FLAVOR. Food Res. 2: 143–150, illus.

(11) ARY, J. E., and JORDAN, R.
1945. PLAIN BUTTER CAKES AND BAKED CUSTARDS MADE FROM SPRAY-DRIED WHOLE-EGG POWDER. Food Res. 10: 476–484, illus.

(12) ASMUNDSON, V. S., JUKES, T. H., FYLER, H. M., and MAXWELL, M. L.
1938. THE EFFECT OF CERTAIN FISH MEALS AND FISH OILS IN THE RATION ON THE FLAVOR OF THE TURKEY. Poultry Sci. 17: 147–151.

(13) BAEDER, H.
1938. THE USE OF LARD IN CAKE MAKING. Nebr. Agr. Expt. Sta. Bul. 320, 19 pp., illus.

(14) BARBELLA, N. G., HANKINS, O. G., and ALEXANDER, L. M.
1936. THE INFLUENCE OF RETARDED GROWTH IN LAMBS ON FLAVOR AND OTHER CHARACTERISTICS OF THE MEAT. Amer. Soc. Anim. Prod. Proc. 29: 289–294.

(15) BATCHELDER, E. L., KIRKPATRICK, M. E., STEIN, K. E., and MARRON, I. M.
1947. EFFECT OF SCALDING METHOD ON QUALITY OF THREE HOME-FROZEN VEGETABLES. Jour. Home Econ. 39: 282–286.

(16) BATEN, W. D.
1945. THE USE OF DISCRIMINANT FUNCTIONS IN COMPARING JUDGES' SCORES CONCERNING POTATOES. Amer. Statis. Assoc. Jour. 40: 223–228.

(17) _____
1946. ORGANOLEPTIC TESTS PERTAINING TO APPLES AND PEARS. Food Res. 11: 84–94, illus.

(18) BATE-SMITH, E. C.
1949. ORGANOLEPTIC TESTS IN THE FOOD INDUSTRY. I. GENERAL INTRODUCTION AND QUALITY OF ANIMAL PRODUCTS. Soc. Chem. Indus. Jour. 68: 78–80, illus.

(19) _____ BROOKS, J., and HAWTHORNE, J. R.
1943. DRIED EGG. PART I. THE PREPARATION, EXAMINATION, AND STORAGE OF SPRAY-DRIED WHOLE EGG. Soc. Chem. Indus. Jour., Trans. and Commun. 62: 97-100.

(20) BEATTIE, G. B.
1939. BEVERAGE OFF-FLAVOURS. Flavours 2 (3): 12, 15–17, illus.

(21) BEATTY, R. M., and CRAGG, L. H.
1935. THE SOURNESS OF ACIDS. Amer. Chem. Soc. Jour. 57: 2347–2351, illus.

(22) BENGTSSON, K., and HELM, E.
1946. PRINCIPLES OF TASTE TESTING. Wallerstein Labs. Commun. 9: 171–180.
(23) BLACK, W. H., HINER, R. L., BURK, L. B., and others.
1940. BEEF PRODUCTION AND QUALITY AS AFFECTED BY METHOD OF FEEDING SUPPLEMENTS TO STEERS ON GRASS IN THE APPALACHIAN REGION. U. S. Dept. Agr. Tech. Bul. 717, 32 pp., illus.
(24) _____ SEMPLE, A. T., and LUSH, J. L.
1934. BEEF PRODUCTION AND QUALITY AS INFLUENCED BY CROSSING BRAHMAN WITH HEREFORD AND SHORTHORN CATTLE. U. S. Dept. Agr. Tech. Bul. 417, 54 pp., illus.
(25) BLAKESLEE, A. F.
1932. GENETICS OF SENSORY THRESHOLDS: TASTE FOR PHENYL-THIO-CARBAMIDE. Natl. Acad. Sci. Proc. 18: 120–130.
(26) _____
1935. A DINNER DEMONSTRATION OF THRESHOLD DIFFERENCES IN TASTE AND SMELL. Science (n. s.) 81: 504–507.
(27) _____
1939. POLL OF 3121 PERSONS SHOWS WIDE VARIATION IN TASTES. Sci. News Letter 35: 51–52.
(28) _____
1942. INDIVIDUALITY AND SCIENCE. Science 95: 1–10.
(29) _____ and SALMON, T. N.
1935. GENETICS OF SENSORY THRESHOLDS: INDIVIDUAL TASTE REACTIONS FOR DIFFERENT SUBSTANCES. Natl. Acad. Sci. Proc. 21: 84–90, illus.
(30) BLANCHARD, E. L., and MAXWELL, M. L.
1941. CORRELATION OF SUBJECTIVE SCORING WITH SUGAR CONTENT OF FROZEN PEAS. Food Res. 6: 105–115, illus.
(31) BLISS, C. I., ANDERSON, E. O., and MARLAND, R. E.
1943. A TECHNIQUE FOR TESTING CONSUMER PREFERENCES, WITH SPECIAL REFERENCE TO THE CONSTITUENTS OF ICE CREAM. Conn. (Storrs Agr. Expt. Sta. Bul. 251, 20 pp.
(32) BOGGS, M. M.
1940. METHODS OF COOKING FROZEN VEGETABLES. West. Canner and Packe 32 (2): 47–48, 54.
(33) _____ CAMPBELL, H., and SCHWARTZE, C. D.
1942. FACTORS INFLUENCING THE TEXTURE OF PEAS PRESERVED BY FREEZING. Food Res. 7: 272–287.
(34) _____ CAMPBELL, H., and SCHWARTZE, C. D.
1943. FACTORS INFLUENCING TEXTURE OF PEAS PRESERVED BY FREEZING. II. Food Res. 8: 502–515.
(35) _____ DUTTON, H. J., EDWARDS, B. G., and FEVOLD, H. L.
1946. DEHYDRATED EGG POWDERS. RELATION OF LIPIDE AND SALT-WATER FLUORESCENCE VALUES TO PALATABILITY. Indus. and Engin. Chem. 38: 1082–1084, illus.
(36) _____ and FEVOLD, H. L.
1946. DEHYDRATED EGG POWDERS. FACTORS IN PALATABILITY OF STORED POWDERS. Indus. and Engin. Chem. 38: 1075–1079, illus.
(37) _____ and HANSON, H. L.
1949. ANALYSIS OF FOODS BY SENSORY DIFFERENCE TESTS. *In* Advances in Food Res., vol. 2, pp. 219–258, illus. New York.
(38) BOHN, R. T.
1935. MILK POWDER FOR THE CAKE BAKING TEST. Cereal Chem. 12: 300–302.
(39) BOHREN, B. B., and JORDAN, R.
1946. AN OBJECTIVE TECHNIQUE FOR DETECTING FLAVOR CHANGES IN DEHYDRATED EGGS. (Abstract) Poultry Sci. 25: 397.
(40) BRADY, D. E.
1937. A STUDY OF THE FACTORS INFLUENCING TENDERNESS AND TEXTURE OF BEEF. Amer. Soc. Anim. Prod. Proc. 30: 246–250.
(41) _____ FREI, P., and HICKMAN, C. W.
1942. EFFECT OF FREEZING RATE ON QUALITY OF BROILED STEAKS. Food Res. 7: 388–393, illus.
(42) BRANAMAN, G. A., HANKINS, O. G., and ALEXANDER, L. M.
1936. THE RELATION OF DEGREE OF FINISH IN CATTLE TO PRODUCTION AND MEAT FLAVORS. Amer. Soc. Anim. Prod. Proc. 29: 295–300.
(43) BRIANT, A. M.
1949. METHODS FOR JUDGING FLAVOR AND ODOR OF COOKED POTATOES FROM SOILS TREATED FOR WIRE WORM CONTROL. Amer. Potato Jour. 26: 300–304.

(44) BRINKMAN, E. V. S., HALLIDAY, E. G., HINMAN, W. F., and HAMNER, R. J.
1942. EFFECT OF VARIOUS COOKING METHODS UPON SUBJECTIVE QUALITIES AND NUTRITIVE VALUES OF VEGETABLES. Food Res. 7: 300–305.
(45) BRISON, F. R.
1945. THE STORAGE OF SHELLED PECANS. Tex. Agr. Expt. Sta. Bul. 667, 16 pp.
(46) BROWN, H. D., MILLER, M. K., ALBAN, K., and others.
1944. CAROTENE, FLAVOR, COLOR AND REFRACTIVE INDICES OF CARROTS GROWN AT DIFFERENT FERTILITY LEVELS. Amer. Soc. Hort. Sci. Proc. 44: 465–467.
(47) BUEL, E. I.
1934. PLAIN CAKE. VIII. THE EFFECT OF THE METHOD OF MIXING ON THE TEXTURE, VOLUME AND PALATABILITY USING LARD AS THE FAT. 111 pp. [Unpublished master's thesis. Copy on file Iowa State Col. Libr., Ames.]
(48) BURTON, W. G.
1949. MASHED POTATO POWDER. IV. DETERIORATION DUE TO OXIDATIVE CHANGES. Soc. Chem. Indus. Jour. 68: 149–151.
(49) CALDWELL, J. S., and CULPEPPER, C. W.
1943. SUITABILITY FOR DEHYDRATION OF 34 VARIETIES AND STRAINS OF SWEET CORN. Canner 96 (19): 12–14, 28; (20): 28, 30–33; (21): 15–16, 28; (22): 20–21, 28, illus.
(50) _____ and CULPEPPER, C. W.
1943. SNAP-BEAN VARIETIES SUITED TO DEHYDRATION. Canning Age 24: 309–313, 363–368, 420–424.
(51) _____ CULPEPPER, C. W., and HUTCHINS, M. C.
1944. VARIETIES OF BEETS SUITED TO DEHYDRATION. Food Packer 25 (9): 47–49; (10): 44–46, 48.
(52) _____ CULPEPPER, C. W., HUTCHINS, M. C., and others.
1944. FURTHER STUDIES OF VARIETAL SUITABILITY FOR DEHYDRATION IN SNAP BEANS. Canner 99 (9): 13–15, 22; (10): 12–15, 30; (11):16–18, 30; (12): 17–18, 38, 40.
(53) _____ CULPEPPER, C. W., HUTCHINS, M. C., and others.
1945. THE DEHYDRATION OF OKRA: VARIETY AND STAGE OF MATURITY AS FACTORS IN DETERMINING QUALITY. Canner 101 (17): 14–16, 22–24, 26.
(54) _____ CULPEPPER, C. W., and SCOTT, D. H.
1943–44. VARIETAL SUITABILITY FOR DEHYDRATION IN EASTERN FREESTONE PEACHES. Fruit Prod. Jour. and Amer. Food Mfr. 23: 68–71, 89, 101–106, 136–142, 151.
(55) _____ CULPEPPER, C. W., and STEVENSON, F. J.
1944. VARIETY, PLACE OF PRODUCTION, AND STAGE OF MATURITY AS FACTORS IN DETERMINING SUITABILITY FOR DEHYDRATION IN WHITE POTATOES. PART II. Canner 99 (1): 26–28, 30, 32, 34; (2): 16, 18, 24, 26, 28, 30.
(56) _____ LOMBARD, P. M., and CULPEPPER, C. W.
1943. VARIETY AND PLACE OF PRODUCTION AS FACTORS IN DETERMINING SUITABILITY FOR DEHYDRATION IN WHITE POTATOES. Canner 97 (3): 30, 32, 34–35, 42, 44; (4): 14–17, 24; (5): 15–16, 18–19, 28.
(57) _____ MOON, H. H., and CULPEPPER, C. W.
1938. A COMPARATIVE STUDY OF SUITABILITY FOR DRYING PURPOSES IN FORTY VARIETIES OF THE SWEETPOTATO. U. S. Dept. Agr. Cir. 499, 52 pp.
(58) CAMERON, A. T.
1944. THE RELATIVE SWEETNESSES OF CERTAIN SUGARS, MIXTURES OF SUGARS, AND GLYCEROL. Canad. Jour. Res. Sect. E, Med. Sci. 22: 45–63, illus.
(59) _____
1947. THE TASTE SENSE AND THE RELATIVE SWEETNESS OF SUGARS AND OTHER SWEET SUBSTANCES. Sugar Res. Found., Sci. Rpt. Ser. 9, 72 pp., illus. New York.
(60) CAMPBELL, H., LINEWEAVER, H., and MORRIS, H. J.
1945. SEVERE BLANCH DOESN'T IMPROVE DEHYDRATED POTATO QUALITY. Food Indus. 17: 384–386, 478, 480, [482], 484, 486, illus.
(61) CARL, B. C., WATTS, B. M., and MORGAN, A. F.
1944. DEHYDRATION OF MEAT SCRAPPLES. Food Res. 9: 319–327.
(62) CARR, R. E., and TROUT, G. M.
1942. SOME COOKING QUALITIES OF HOMOGENIZED MILK. I. BAKED AND SOFT CUSTARD. Food Res. 7: 360–369, illus.

(63) CARTWRIGHT, L. C., and NANZ, R. A.
1948. ORGANOLEPTIC EVALUATION OF FOODS. (Abstract) Spice Mill 71 (8): 47–48, 51.
(64) _____ and NANZ, R. A.
1948. FLAVORS IMPROVED, SALES BOOSTED THROUGH ORGANOLEPTIC TESTS. Food Indus. 20: 1608–1611, 1710–1712, illus.
(65) _____ SNELL, C. T., and KELLEY, P. H.
1949. ORGANOLEPTIC PANEL TESTING AS A RESEARCH TOOL. 15 pp. (Foster D. Snell, Inc.) New York. [Processed.]
(66) CATHCART, W. H.
1937. EXPERIMENTS IN DETERMINING BREAD FLAVOR. Cereal Chem. 14: 735–751.
(67) _____ and KILLEN, E. J.
1940. SCORING OF TOAST AND FACTORS WHICH AFFECT ITS QUALITY. Food Res. 5: 307–321.
(68) CLARK, R. K., and VAN DUYNE, F. O.
1949. COOKING LOSSES, TENDERNESS, PALATABILITY, AND THIAMINE AND RIBOFLAVIN CONTENT OF BEEF AS AFFECTED BY ROASTING, PRESSURE SAUCEPAN COOKING, AND BROILING. Food Res. 14: 221–230. (Portion of thesis, M. S., Univ. Ill.)
(69) CLENDENNING, T.
1940. FLAVOR IN CONFECTIONS. PART II—METHODS OF EVALUATION. Mfg. Confectioner 20 (2): 23–25.
(70) CLINE, J. A., TROWBRIDGE, E. A., FOSTER, M. T., and FRY, H. E.
1930. HOW CERTAIN METHODS OF COOKING AFFECT THE QUALITY AND PALATABILITY OF BEEF. Mo. Agr. Expt. Sta. Bul. 293, 40 pp., illus.
(71) COBB, J. S.
1935. A STUDY OF CULINARY QUALITY IN WHITE POTATOES. Amer. Potato Jour. 12: 335–346.
(72) COLE, W. C., and BOULWARE, J. H.
1940. INFLUENCE OF SOME MIX COMPONENTS UPON THE TEXTURE OF ICE CREAM. Jour. Dairy Sci. 23: 149–157, illus.
(73) CONRAD, R. M., VAIL, G. E., OLSEN, A. L., and others.
1948. IMPROVED DRIED WHOLE EGG PRODUCTS. Kans. Agr. Expt. Sta. Tech. Bul. 64, 62 pp., illus.
(74) CORBETT, W. J., and TRACY, P. H.
1941. EXPERIMENTS ON THE USE OF CERTAIN ANTIOXIDANTS FOR CONTROL OF OXIDIZED FLAVOR IN DAIRY PRODUCTS. Food Res. 6: 445–459.
(75) COVER, S.
1936. A NEW SUBJECTIVE METHOD OF TESTING TENDERNESS IN MEAT — THE PAIRED-EATING METHOD. Food Res. 1: 287–295.
(76) _____
1937. THE EFFECT OF TEMPERATURE AND TIME OF COOKING ON THE TENDERNESS OF ROASTS. Tex. Agr. Expt. Sta. Bul. 542, 61 pp., illus.
(77) _____
1940. SOME MODIFICATIONS OF THE PAIRED-EATING METHOD IN MEAT COOKERY RESEARCH. Food Res. 5: 379–394.
(78) _____
1941. EFFECT OF METAL SKEWERS ON COOKING TIME AND TENDERNESS OF BEEF. Food Res. 6: 233–238, illus.
(79) _____
1943. EFFECT OF EXTREMELY LOW RATES OF HEAT PENETRATION ON TENDERING OF BEEF. Food Res. 8: 388–394.
(80) COWAN, J. C., and CLOPTON, J. R.
1945. ORGANOLEPTIC PROPERTIES OF SOYBEAN OIL. *In* Deterioration of Fats and Oils, Quartermaster Corps Manual QMC 17–7: 77–83.
(81) COX, G. M.
1944. STATISTICS AS A TOOL FOR RESEARCH. Jour. Home Econ. 36: 575–580.
(82) CRIST, J. W., and SEATON, H. L.
1941. RELIABILITY OF ORGANOLEPTIC TESTS. Food Res. 6: 529–536.
(83) CROCKER, E. C.
1935. THE NATURE OF FLAVOR. U. S. Egg and Poultry Mag. 41 (11): 14–15, 63–64.
(84) _____
1937. MEASURING FOOD FLAVORS. Food Res. 2: 273–286, illus.
(85) _____
1945. FLAVOR. 172 pp., illus. New York and London.
(86) _____
1947. ODOR IN FLAVOR. Spice Mill 70 (10): 69–70, illus.

(87) _____
1948. FLAVOR OF MEAT. Food Res. 13: 179–183.
(88) _____ and SJÖSTRÖM, L. B.
1949. ODOR DETECTION AND THRESHOLDS. Chem. and Engin. News 27: 1922–1925.
(89) _____ SJÖSTRÖM, L. B., and TALLMAN, G. B.
1948. MEASUREMENT OF FOOD ACCEPTANCE. Indus. and Engin. Chem. 40: 2254–2257.
(90) CRUICKSHANK, E. M.
1939. THE EFFECT OF COD-LIVER OIL AND FISHMEAL ON THE FLAVOR OF POULTRY PRODUCTS. *In* Seventh World's Poultry Cong. and Expo. Proc., pp. 539–542, illus.
(91) CULPEPPER, C. W., CALDWELL, J. S., HUTCHINS, M. C., and others.
1945. DEHYDRATION OF PUMPKIN AND WINTER SQUASH. A COMPARATIVE STUDY OF FORTY-TWO VARIETIES. Fruit Prod. Jour. and Amer. Food Mfr. 24: 170–177, 189, 202–208, 215.
(92) DAHLBERG, A. C.
1945. THE KEEPING QUALITY OF PASTEURIZED MILK IN THE NEW YORK METROPOLITAN AREA DURING COOL WEATHER AS DETERMINED BY BACTERIAL COUNTS, PRESENCE OF COLIFORM BACTERIA, AND FLAVOR SCORES. Jour. Dairy Sci. 28: 779–792.
(93) _____ and PENCZEK, E. S.
1941. THE RELATIVE SWEETNESS OF SUGARS AS AFFECTED BY CONCENTRATION. N. Y. State Agr. Expt. Sta. Tech. Bul. 258, 12 pp., illus.
(94) DAVIS, M. V., and HALLIDAY, E. G.
1939. STUDIES ON ALL-PURPOSE FLOUR. Cereal Chem. 16: 414–418.
(95) DAWSON, E. H., compiler
1943. EXPERIMENTAL PROCEDURE FOR CONDUCTING TASTE AND SMELL TESTS. 8 pp. (U. S. Bureau of Human Nutrition and Home Economics.) [Processed.]
(96) _____ DUEHRING, M., and PARKS, V. E.
1947. ADDITION OF GROUND EGG SHELL TO DRIED EGG FOR USE IN COOKING. Food Res. 12: 288–297.
(97) _____ REYNOLDS, H., and TOEPFER, E. W.
1949. HOME-CANNED VERSUS HOME-FROZEN SNAP BEANS. Jour. Home Econ. 41: 572–574.
(98) _____ SHANK, D. E., LYNN, J. M., and WOOD, E. A.
1945. EFFECT OF STORAGE . . . ON FLAVOR AND COOKING QUALITY OF SPRAY-DRIED WHOLE EGG. U. S. Egg and Poultry Mag. 51: 154–161, illus.
(99) _____ WOOD, E. A., and MCNALLY, E. H.
1947. SPRAY DRIED WHOLE EGGS IMPROVED WITH CARBOHYDRATES. Food Indus. 19: 483–486, 594–596, illus.
(100) DERMER, O. C.
1947. THE SCIENCE OF TASTE. Okla. Acad. Sci. Proc. 27: [9] 20.
(101) DOVE, W. F.
1943. THE RELATIVE NATURE OF HUMAN PREFERENCE: WITH AN EXAMPLE IN THE PALATABILITY OF DIFFERENT VARIETIES OF SWEET CORN. Jour. Compar. Psychol. 35: 219–226.
(102) _____
1946. DEVELOPING FOOD ACCEPTANCE RESEARCH. Science 103: 187–190.
(103) _____
1947. FOOD ACCEPTABILITY — ITS DETERMINATION AND EVALUATION. Food Technol. 1: 39–50.
(104) _____ and FARRELL, B. L.
1945. TECHNICS FOR MEASURING CHANGES IN FLAVOR ACCEPTABILITY. *In* Deterioration of Fats and Oils, Quartermaster Corps Manual QMC 17–7: 84–88, illus.
(105) DOWNS, P. A.
1937. JUDGING QUALITY IN DAIRY PRODUCTS. Nebr. Agr. Expt. Sta. Cir. 54, 44 pp., illus.
(106) DUBOIS, C. W., TRESSLER, D. K., and FENTON, F.
1940. INFLUENCE OF RATE OF FREEZING AND TEMPERATURE OF STORAGE ON QUALITY OF FROZEN MEAT. *In* Inst. Food Technol. Proc., pp. 167–179, illus.
(107) _____ TRESSLER, D. K., and FENTON, F.
1942. THE EFFECT OF THE RATE OF FREEZING AND TEMPERATURE OF STORAGE ON THE QUALITY OF FROZEN POULTRY. Refrig. Engin. 44: 93–99, 122, illus.

(108) DUNKER, C. F., TRESSLER, D. K., WRUCK, M., and BLAKE, K. B.
1947. RELATIONSHIP OF MOISTURE CONTENT TO QUALITY RETENTION IN DEHYDRATED VEGETABLES DURING STORAGE. I. — TOMATO FLAKE. Food Technol. 1: 17–25, illus.
(109) ELSBERG, C. A., BREWER, E. D., and LEVY, I.
1935. THE SENSE OF SMELL. VII. THE ODOROUS SUBSTANCES TO BE USED FOR TESTS OF THE OLFACTORY SENSE. Bul. Neurological Inst. New York 4: 286–293, illus.
(110) _____ LEVY, I., and BREWER, E. D.
1936. A NEW METHOD FOR TESTING THE SENSE OF SMELL AND FOR THE ESTABLISHMENT OF OLFACTORY VALUES OF ODOROUS SUBSTANCES. Science (n. s.) 83: 211–212.
(111) FABIAN, F. W.
1940. WHAT IS FLAVOR? Canner 91 (2): 12–13, 20; (3): 13–15, illus.
(112) _____ and BLUM, H. B.
1943. RELATIVE TASTE POTENCY OF SOME BASIC FOOD CONSTITUENTS AND THEIR COMPETITIVE AND COMPENSATORY ACTION. Food Res. 8: 179–193.
(113) FAULKNER, E. C., and SIMPSON, J. I.
1946. A COMPARATIVE STUDY OF SOY FLOURS FOR USE IN BAKING. II. EFFECT ON QUALITY OF MUFFINS AND CAKES. Food Res. 11: 454–459.
(114) FENTON, F.
1946. INFLUENCE OF METHOD OF PREPARATION ON RETENTION OF PALATABILITY AND VITAMINS IN FOOD. *In* Food Acceptance Research, Quartermaster Corps Manual QMC 17–9: 20–26.
(115) _____ and GIFFT, H.
1943. PALATABILITY STUDIES OF COMMERCIALLY DEHYDRATED VEGETABLES. I. EFFECT OF SEVERAL METHODS OF STORAGE ON PALATABILITY OF BEETS, CABBAGE, AND RUTABAGAS. II. EFFECT OF SEVERAL COMMON REFRESHING AND COOKING METHODS ON PALATABILITY AND WATER ABSORPTION OF BEETS, CABBAGE, POTATOES, RUTABAGAS, AND YELLOW TURNIPS. Food Res. 8: 364–376.
(116) FEVOLD, H. L., EDWARDS, B. G., DIMICK, A. L., and BOGGS, M. M.
1946. DEHYDRATED EGG POWDERS. SOURCES OF OFF-FLAVORS DEVELOPED DURING STORAGE. Indus. and Engin. Chem. 38: 1079–1082.
(117) FITZGERALD, G. A., and NICKERSON, J. T. R.
1939. EFFECT OF TIME AND TEMPERATURE OF HOLDING UNDRAWN POULTRY UPON ITS QUALITY. *In* Seventh World's Poultry Cong. and Expo. Proc., pp. 509–512.
(118) FORD, L. A.
1940. THE NOSE IN THE CHEMISTRY LABORATORY. Jour. Chem. Ed. 17: 17–19.
(119) FORSTER, T. L., and BROWN, R. W.
1943. COMPARATIVE STUDIES IN THE PREVENTION OF WOOD TAINT, TOPPINESS AND PRIMROSE COLOUR OF STORAGE BUTTER. Sci. Agr. 23: 342–354.
(120) FREEMAN, M. E.
1941. SCORING BAKED POTATOES FOR TEXTURE. Food Res. 6: 595–598, illus.
(121) FRYD, C. F. M., and HANSON, S. W. F.
1945. DRIED EGG. THE RELATIONSHIP OF ANALYSIS TO FLAVOUR. Soc. Chem. Indus. Jour., Trans. and Commun. 64: 55–56, illus.
(122) GAEBE, O. F.
1940. A COMPARATIVE ODOR AND FLAVOR STUDY OF EGGS STORED IN AVENIZED AND UNAVENIZED FILLERS AND FLATS. U. S. Egg and Poultry Mag. 46: 346–349, illus.
(123) GARDNER, B. W., JR.
1949. ARMY TESTS REVEAL HOW STORAGE AFFECTS CANNED MEAT FLAVOR. Food Indus. 21: 889–890, illus.
(124) GELMAN, G.
1945. PSYCHOMETRICS — NEW QUALITY CONTROL TOOL? Food Indus. 17: 625, 720, 722, 724.
(125) GILMER, R. S., KINDER, D. E., and BOHN, R. M.
1936. THE DISAPPEARANCE OF FLAVORS IN BISCUITS. Cereal Chem. 13: 421–427.
(126) GLEIM, E., and FENTON, F.
1949. EFFECT OF 0° F. AND 15° F. STORAGE ON THE QUALITY OF FROZEN COOKED FOODS. Food Technol. 3: 187–192, illus.
(127) GORTNER, W. A., FENTON, F., VOLZ, F. E., and GLEIM, E.
1948. EFFECT OF FLUCTUATING STORAGE TEMPERATURES ON QUALITY OF FROZEN FOODS. Indus. and Engin. Chem. 40: 1423–1426.

(128) GRANT, G. A., and LIPS, H. J.
1946. A STUDY OF METHODS FOR ASSESSING RANCIDITY IN LARD. Canad. Jour. Res. Sect. F, Technol. 24: 450–460, illus.
(129) _____ and WHITE, W. H.
1946. A FLUORESCENCE METHOD FOR ASSESSING THE KEEPING QUALITY OF BUTTER. Canad. Jour. Res. Sect. F, Technol. 24: 461–466, illus.
(130) GRAUL, L. S., and LOWE, B.
1947. HOW STORAGE AFFECTS FROZEN CAKES AND BATTERS. Food Indus. 19: 330–332, illus.
(131) GREEN, G. H.
1949. ORGANOLEPTIC TESTS IN THE FOOD INDUSTRY. (Letter to the editor.) Chem. and Indus. no. 32, p. 567.
(132) GREENBANK, G. R., WRIGHT, P. A., DEYSHER, E. F., and HOLM, G. E.
1946. THE KEEPING QUALITY OF SAMPLES OF COMMERCIALLY DRIED MILK PACKED IN AIR AND IN INERT GAS. Jour. Dairy Sci. 29: 55–61, illus.
(133) GREENWOOD, M. L., and SALERNO, R.
1947. PALATABILITY OF KALE IN RELATION TO VARIETY AND COOKING PROCEDURES. Natl. Coop. Proj. Conservation of Nutritive Value of Foods, Prog. Notes 2, 3 pp. (Conn. (Storrs) Agr. Expt. Sta.) [Processed.]
(134) _____ and TICE, J. M.
1949. PALATABILITY TESTS ON POTATOES GROWN IN SOIL TREATED WITH THE INSECTICIDES BENZENE HEXACHLORIDE, CHLORDANE, AND CHLORINATED CAMPHENE. Jour. Agr. Res. 78: 477–482.
(135) GRISWOLD, R. M.
1940. PALATABILITY AND COLOR OF POTATOES BOUGHT ON A RETAIL MARKET. Food Res. 5: 281–290.
(136) _____
1944. FACTORS INFLUENCING THE QUALITY OF HOME-CANNED MONTMORENCY CHERRIES. Mich. Agr. Expt. Sta. Tech. Bul. 194, 38 pp., illus.
(137) _____
1944. FACTORS INFLUENCING THE QUALITY OF COOKED JONATHAN APPLES. Mich. Agr. Expt. Sta. Tech. Bul. 195, 19 pp.
(138) _____ and WHARTON, M. A.
1941. EFFECT OF STORAGE CONDITIONS ON PALATABILITY OF BEEF. Food Res. 6: 517–528.
(139) GROVER, D. W., and HAWTHORNE, J. R.
1946. AN INVESTIGATION OF THE CHARACTERS OF DRIED WHOLE EGG DETERMINING BAKING QUALITY FOR CAKES OTHER THAN SPONGE CAKES. Food Res. 11: 41–48.
(140) HANDSCHUMAKER, E.
1948. A TECHNIQUE FOR TESTING THE REVERSION PROPERTIES OF HYDROGENATED SOYBEAN OIL SHORTENINGS. Amer. Oil Chem. Soc. Jour. 25: 54–56.
(141) HANSON, H. L., LOWE, B., and STEWART, G. F.
1947. PASTEURIZATION OF LIQUID EGG PRODUCTS. V. THE EFFECT ON PERFORMANCE IN CUSTARDS AND SPONGE CAKES. Poultry Sci. 26: 277–283, illus.
(142) _____ STEWART, G. F., and LOWE, B.
1942. PALATABILITY AND HISTOLOGICAL CHANGES OCCURRING IN NEW YORK DRESSED BROILERS HELD AT 1.7°C. (35°F.). Food Res. 7: 148–160, illus.
(143) HARDING, P. L.
1947. QUALITY IN CITRUS FRUITS. Citrus Indus. 28 (2): 5–9, 20, illus.
(144) _____ and WADLEY, F. M.
1945. STUDY OF QUALITY IN TEMPLE ORANGES. Food Res. 10: 510–517, illus.
(145) _____ and WADLEY, F. M.
1948. TEEN-AGE STUDENTS VERSUS ADULTS AS TASTE JUDGES OF TEMPLE ORANGES. Food Res. 13: 6–10.
(146) HARDING, T. S.
1948. TEACHING TEST TASTERS TO TASTE. Pract. Home Econ. 26: 105, illus.
(147) HARDY, F., and NOBLE, I.
1945. A COMPARISON OF MEASUREMENT OF JUICINESS IN ROAST PORK LOIN BY PRESS-FLUID AND JURY-RATING METHODS. Food Res. 10: 160–164.
(148) HARPER, R.
1949. FOOD GRADING AND ITS STUDY. Food 18: 207–210.
(149) HARRIS, R. H., and KNOWLES, D.
1943. MACARONI COOKING VALUE OF SOME NORTH DAKOTA DURUM WHEAT SAMPLES. Food Res. 8: 292–298, illus.

(150) HARSHAW, H. M., HALE, W. S., SWENSON, T. L., and others.
1941. QUALITY OF FROZEN POULTRY AS AFFECTED BY STORAGE AND OTHER CONDITIONS. U. S. Dept. Agr. Tech. Bul. 768, 20 pp.
(151) HELM, E., and TROLLE, B.
1946. SELECTION OF A TASTE PANEL. Wallerstein Labs. Commun. 9: 181–194, illus.
(152) HENING, J. C.
1948. FLAVOR EVALUATION PROCEDURES. N. Y. State Agr. Expt. Sta. Tech. Bul. 284, 20 pp., illus.
(153) _____
1949. OPERATION OF A ROUTINE TESTING GROUP IN A SMALL LABORATORY. Food Technol. 3: 162–163.
(154) HICKS, E. W.
1948. TASTING TESTS. Food Preserv. Quart. 8 (1): 1–5.
(155) HLYNKA, I., HOOD, E. G., and GIBSON, C. A.
1943. AGITATION AND TEMPERATURE OF CHEESE MILK AND THE DEVELOPMENT OF RANCID AND UNCLEAN FLAVORS IN CHEDDAR CHEESE. Jour. Dairy Sci. 26: 1111–1119.
(156) HOPKINS, J. W.
1946. PRECISION OF ASSESSMENT OF PALATABILITY OF FOODSTUFFS BY LABORATORY PANELS. Canad. Jour. Res. Sect. F, Technol. 24: 203–214.
(157) HOTALING, N., and FENTON, F.
1945. CEREAL-EXTENDED GROUND PORK MEAT LOAVES: THEIR PALATABILITY. Jour. Home Econ. 37: 629–638.
(158) HOWE, P. E., and BARBELLA, N. G.
1937. THE FLAVOR OF MEAT AND MEAT PRODUCTS. Food Res. 2: 197–202.
(159) INGELS, B. D., IRWIN, R., and LANDIS, Q.
1936. REPORT OF THE BREAD JUDGING COMMITTEE. Cereal Chem. 13: 218–221.
(160) JACK, E. L., TARASSUK, N. P., and SCARAMELLA, E. L.
1942. EFFECT OF DIFFERENT PASTEURIZATION TEMPERATURES ON KEEPING QUALITY OF BUTTER MADE FROM CREAM CONTAINING NATURALLY ACTIVE LIPASE. Natl. Butter and Cheese Jour. 33 (12): 16–18, 20.
(161) JACOBS, M. B.
1948. FLAVOR MEASUREMENT. PROPER MEASUREMENT OF FLAVOR IS THE KEY TO ADEQUATE APPRAISAL OF FLAVOR ACCEPTABILITY. Amer. Perfumer and Essential Oil Rev. 52: 133, 135. (Reprinted in Flavours, Fruit Juices and Spices Rev. 11 (3): 12–14 [1949.])
(162) JAKOBSEN, F.
1949. RATIONAL GRADING OF FOOD QUALITY. Food Technol. 3: 252–254, illus.
(163) JORDAN, R., and SISSON, M. S.
1943. USE OF SPRAY-DRIED WHOLE EGGS IN BAKED CUSTARDS. U. S. Egg and Poultry Mag. 49: 266–269, 287–288.
(164) JOSEPHSON, D. V.
1945–46. PSYCHOMETRIC AND ORGANOLEPTIC PROCEDURES FOR DETERMINATION OF FLAVOR VALUES IN SELECTED DAIRY PRODUCTS. Com. on Food Res., Proj. Rpt. 1, 3 pp.; Proj. Rpt. 2, 4 pp.; Proj. Rpt. 3, 3 pp. (Office of the Quartermaster General.) [Processed.]
(165) JOSLYN, M. A., JONES, W. L., LAMBERT, E., and others.
1949. APPLICATIONS OF SUGAR SOLUTIONS WITH AND WITHOUT ANTIOXIDANTS. PART IV. ACCEPTABILITY APPRAISAL (SUBJECTIVE QUALITY APPRAISAL). Quick Frozen Foods and Locker Plant 12 (3): 49–52, 86.
(166) KING, F. B.
1937. OBTAINING A PANEL FOR JUDGING FLAVOR IN FOODS. Food Res. 2: 207–219, illus.
(167) _____ COLEMAN, D. A., and LECLERC, J. A.
1937. REPORT OF THE U. S. DEPARTMENT OF AGRICULTURE BREAD FLAVOR COMMITTEE. Cereal Chem. 14: 49–58.
(168) _____ LOUGHLIN, R., RIEMENSHNEIDER, R. W., and ELLIS, N. R.
1936. THE RELATIVE VALUE OF VARIOUS LARDS AND OTHER FATS FOR THE DEEP-FAT FRYING OF POTATO CHIPS. Jour. Agr. Res. 53: 369–381, illus.
(169) KNOWLES, D., and JOHNSON, P. E.
1941. A STUDY OF THE SENSITIVENESS OF PROSPECTIVE FOOD JUDGES TO THE PRIMARY TASTES. Food Res. 6: 207–216.
(170) KOHLMEYER, W., and SHAFFNER, C. S.
1944. COMPARATIVE QUALITY LOSSES IN EGGS STORED IN FIBRE AND WOODEN CASES. U. S. Egg and Poultry Mag. 50: 295–297, 336.

(171) KOONZ, C. H., and TRELEASE, R. D.
1946. ORGANOLEPTIC SIGNIFICANCE OF KIDNEYS IN POULTRY. Food Res. 11: 542–545.
(172) KRAMER, A., and MAHONEY, C. H.
1940. COMPARISON OF ORGANOLEPTIC AND PHYSICOCHEMICAL METHODS FOR DETERMINING QUALITY IN FRESH, FROZEN, AND CANNED LIMA BEANS. Food Res. 5: 583–592.
(173) LAIRD, D. A., and BREEN, W. J.
1939. SEX AND AGE ALTERATIONS IN TASTE PREFERENCES. Amer. Dietet. Assoc. Jour. 15: 549–550, illus.
(174) LAMPITT, L. H., and MORAN, T.
1933. THE PALATABILITY OF RAPIDLY FROZEN MEAT. Soc. Chem. Indus. Jour., Trans. and Commun. 52: 143T–146T, illus.
(175) LANGWILL, K. E.
1949. TASTE PERCEPTION AND TASTE PREFERENCE OF THE CONSUMER. Food Technol. 3: 136–139, illus.
(176) LEE, B. F.
1934. A GENETIC ANALYSIS OF TASTE DEFICIENCY IN THE AMERICAN NEGRO. Ohio Jour. Sci. 34: 337–342.
(177) LEE, F. A., GORTNER, W. A., and WHITCOMBE, J.
1946. EFFECT OF FREEZING RATE ON VEGETABLES. APPEARANCE, PALATABILITY, AND VITAMIN CONTENT OF PEAS AND SNAP BEANS. Indus. and Engin. Chem. 38: 341–346, illus.
(178) LEIGHTON, A., and WILLIAMS, O. E.
1943. SWEETENING POWER OF THE CORN SUGARS IN ICE CREAM. Jour. Dairy Sci. 26: 1107–1110.
(179) LEMON, H. W., LIPS, A., and WHITE, W. H.
1945. FLAVOUR REVERSION IN HYDROGENATED LINSEED OIL. II. EFFECT OF VARIATIONS IN PROCESSING PROCEDURES. Canad. Jour. Res. Sect. F, Technol. 23: 295–303.
(180) LEVIN, G.
1943. TASTE SCORING TESTS ON DRIED WHOLE EGGS. U. S. Egg and Poultry Mag. 49: 371, 375–377.
(181) LIGHTBODY, H. D.
[n. d.] SUMMARY OF INVESTIGATIONS ON BIOCHEMISTRY OF DETERIORATIVE CHANGES IN DEHYDRATED EGGS. *In* Off. Quartermaster General, Subsist. Res. and Devlpmt. Lab. Proc., Conf. 1, pp. 50–55.
(182) LORANT, J., and LORANT, G.
1948. TASTE TEST PROCEDURE. 3 pp. (Oreg. Agr. Expt. Sta., Food Technol. Dept.) [Processed.]
(183) LOWE, B.
1939. EFFECT OF DRAWING BEFORE FREEZING ON THE PALATABILITY OF POULTRY. *In* Seventh World's Poultry Cong. and Expo. Proc., pp. 500–505, illus.
(184) _____
1948. FACTORS AFFECTING THE PALATABILITY OF POULTRY WITH EMPHASIS ON HISTOLOGICAL POST-MORTEM CHANGES. *In* Advances in Food Res., vol. 1, pp. 203–256, illus. New York.
(185) _____ and STEWART, G. F.
1947. SUBJECTIVE AND OBJECTIVE TESTS AS FOOD RESEARCH TOOLS WITH SPECIAL REFERENCE TO POULTRY MEAT. Food Technol. 1: 30–38, illus.
(186) LOY, H. W., HALL, J. L., MACKINTOSH, D. L., and others.
1944. QUALITY OF BEEF. PART I. MINERAL CONSTITUENTS OF BLOOD, MUSCLE TISSUE, AND FAT TISSUE OF BEEF ANIMALS AND THEIR RELATION TO KEEPING QUALITY. PART II. EFFECT OF DIETARY PHOSPHORUS DEFICIENCY ON QUALITY OF BEEF. PART III. EFFECT OF FEEDING LIMESTONE SUPPLEMENTS ON QUALITY OF BEEF. PART IV. CHARACTERISTICS OF DARK-CUTTING BEEF. SURVEY AND PRELIMINARY INVESTIGATION. Kans. Agr. Expt. Sta. Tech. Bul. 58, 86 pp., illus.
(187) MCCAMMON, R. B., PITTMAN, M. S., and WILHELM, L. A.
1934. THE ODOR AND FLAVOR OF EGGS. Poultry Sci. 13: 95–101.
(188) MCCARTHY, I.
[n. d.] HOW CAN WE MAKE SURE OF PLEASING THE TASTES OF MRS. AMERICA AND HER FAMILY? 6 pp. (Hallmark Testing Service, Safeway Stores, Inc., Oakland, Calif.) [Processed.]
(189) MCINTOSH, J. A., TANNER, R., EVANS, R. J., and CARVER, J. S.
1942. COOKING PROPERTIES OF EGGS PROCESSED IN MINERAL OIL. U. S. Egg and Poultry Mag. 48: 345–347, 383.

(190) McKim, E., and Moss, H. V.
1939. STUDY OF DEFINITION AND EVALUATION OF VARIOUS ITEMS ON SCORE CARD. Cereal Chem. 16: 117–126, illus.
(191) Maiden, A. M.
1936. A SYSTEM OF JUDGING FLAVOUR IN BREAD. Chem. and Indus., no. 55, pp. 143–145.
(192) Marble, D. R., Hunter, J. E., Knandel, H. C., and Dutcher, R. A.
1938. FISHY FLAVOR AND ODOR IN TURKEY MEAT. Poultry Sci. 17: 49–53.
(193) Marcuse, S.
1945. AN APPLICATION OF THE CONTROL CHART METHOD TO THE TESTING AND MARKETING OF FOODS. Amer. Statis. Assoc. Jour. 40: 214–222, illus.
(194) ———
1947. APPLYING CONTROL CHART METHODS TO TASTE TESTING. Food Indus. 19: 316–318, illus.
(195) Marion, P. T., Butler, O. D., Cover, S., and Jones, J. H.
1948. RATE OF GAIN AND TENDERNESS OF BEEF. Tex. Agr. Expt. Sta. Prog. Rpt. 1125, Cattle Ser. 70, 3 pp. [Processed.]
(196) Marshall, J. B., Grant, G. A., and White, W. H.
1945. RATION BISCUITS. III. EFFECT OF MOISTURE CONTENT ON KEEPING QUALITY. Canad. Jour. Res. Sect. F, Technol. 23: 286–294, illus.
(197) Martin, E. L.
1933. THE CREAMING QUALITIES OF LARDS AND OTHER FATS AND NEW WAYS OF MIXING THEM IN PLAIN CAKES. 151 pp., illus. [Unpublished master's thesis. Copy on file Iowa State Col. Libr., Ames.]
(198) Martin, W. H., Nelson, F. E., and Caulfield, W. J.
1940. MEASURING THE QUALITY OF ICE CREAM. Jour. Dairy Sci. 23: 135–147.
(199) Mason, H. M., and Lipscomb, A. G.
1948. OBJECTIVE TASTING TESTS. Chem. and Indus., no. 7, p. 107.
(200) Maw, W. A.
1935. HOW QUALITY IN POULTRY MEAT IS AFFECTED BY THE DISTRIBUTION OF FAT IN THE CARCASS. A STUDY OF NUTRITIONAL FACTORS INFLUENCING THE QUALITY AS CONSUMERS FIND IT. U. S. Egg and Poultry Mag. 41 (5): 33–36, illus.
(201) Maybee, G. R.
1939. FLAVOUR IN FOOD. CLASSIFICATION AND COMPARISON OF TASTES. (Abstract) Canad. Chem. and Process Indus. 23: 115–118.
(202) Metzner, C. A.
1943. INVESTIGATION OF ODOR AND TASTE. PSYCHOLOGICAL PRINCIPLES. Wallerstein Labs. Commun. 6: 5–18, illus.
(203) Meyer, B., Buckley, R., and Moore, R.
1949. RESEARCH SHOWS DIFFERENCES IN FROZEN BUTTER CAKE AND SPONGE CAKE. Refrig. Engin. 57: 340–342, 388, 392, illus.
(204) Miller, C., and Beattie, I. E.
1949. ON THE FREEZING AND FROZEN STORAGE OF CAKE. Jour. Home Econ. 41: 463–464.
(205) Miller, C. F., Lowe, B., and Stewart, G. F.
1947. LIFTING POWER OF DRIED WHOLE EGG WHEN USED IN SPONGE CAKE. Food Res. 12: 332–342, illus.
(206) Moir, H. C.
1936. SOME OBSERVATIONS ON THE APPRECIATION OF FLAVOUR IN FOODSTUFFS. Chem. and Indus., no. 55, pp. 145–148.
(207) ———
1947. ESTIMATION AND CONTROL OF FLAVOUR IN FOODSTUFFS. Perfumery and Essential Oil Rec. 38: 299–301.
(208) Moncrieff, R. W.
1944. GUSTATION. PART III. Food Manufacture 19: 356–361, illus.
(209) ———
1946. TESTING THE FLAVOUR OF DRIED EGGS. Food Manufacture 21: 55–58, 63, illus.
(210) ———
1947. THE CHOICE OF FLAVOUR. Food Manufacture 22: 113–118.
(211) ———
1948. ODOUR AND OXIDISABILITY. Food Manufacture 23: 411–413.
(212) ———
[1949.] RELATIVE SWEETNESS. PART I. Flavours, Fruit Juices and Spices Rev. 11 (5): 5–6, 8.
(213) Moser, H. A., Jaeger, C. M., Cowan, J. C., and Dutton, H. J.
1947. THE FLAVOR PROBLEM OF SOYBEAN OIL. II. ORGANOLEPTIC EVALUATION. Amer. Oil Chem. Soc. Jour. 24: 291–296, illus.

(214) MURPHY, R. R., BOUCHER, R. V., and KNANDEL, H. C.
1939. FLAVOR OF TURKEY MEAT AS AFFECTED BY FEEDING FISHMEAL AND FISH OIL. *In* Seventh World's Poultry Cong. and Expo. Proc., pp. 542–545.
(215) NELSON, P. M., LOWE, B., and HELSER, M. D.
1930. INFLUENCE OF THE ANIMAL'S AGE UPON THE QUALITY AND PALATABILITY OF BEEF. PART II. THE ROAST BEEF, PREPARATION, QUALITY AND PALATABILITY. *In* Iowa Agr. Expt. Sta. Bul. 272: 311–323, illus.
(216) NOBLE, I., and HARDY, F.
1945. EFFECT OF STORAGE TEMPERATURE AND TIME UPON QUALITY OF PORK PRESERVED BY FREEZING. Food Res. 10: 165–175, illus.
(217) OVERMAN, A., and LI, J. C. R.
1948. DEPENDABILITY OF FOOD JUDGES AS INDICATED BY AN ANALYSIS OF SCORES OF A FOOD-TASTING PANEL. Food Res. 13: 441–449.
(218) PAUL, P., FRUEH, M., and OHLSON, M. A.
1948. CORN MEAL AND MACARONI PRODUCTS CONTAINING DRY PRIMARY YEAST. I. PALATABILITY AND ACCEPTABILITY. Amer. Dietet. Assoc. Jour. 24: 673–675.
(219) _____ LOWE, B., and MCCLURG, B. R.
1944. CHANGES IN HISTOLOGICAL STRUCTURE AND PALATABILITY OF BEEF DURING STORAGE. Food Res. 9: 221–233, illus.
(220) _____ and MCLEAN, B. B.
1946. STUDIES ON VEAL. I. EFFECT OF DIFFERENT INTERNAL TEMPERATURES ON VEAL ROASTS FROM CALVES OF THREE DIFFERENT WEIGHTS. Food Res. 11: 107–115, illus.
(221) PEARCE, J. A.
1943. THE DEHYDRATION OF PORK. Canad. Jour. Res. Sect. D, Zool. Sci. 21: 394–404, illus.
(222) _____
1944 FLUORESCENCE DEVELOPMENT IN VARIOUS FOOD PRODUCTS. Canad. Jour. Res. Sect. F, Technol. 22: 87–95, illus.
(223) _____
1945. FACTORS AFFECTING THE STORAGE OF DEHYDRATED PORK. Canad. Jour. Res. Sect. F, Technol. 23: 9–21, illus.
(224) _____
1945. DRIED MILK POWDER. I. METHODS OF ASSESSING QUALITY AND SOME EFFECTS OF HEAT TREATMENT. Canad. Jour. Res. Sect. F, Technol. 23: 177–184, illus.
(225) _____ and BRYCE, W. A.
1945. DRIED MILK POWDER. III. THE EFFECT OF LIGHT ON KEEPING QUALITY. Canad. Jour. Res. Sect. F, Technol. 23: 334–339, illus.
(226) _____ WHITTAKER, J., TESSIER, H., and BRYCE, W. A.
1946. THE KEEPING QUALITY OF DEHYDRATED MIXTURES OF EGG AND MILK. Canad. Jour. Res. Sect. F, Technol. 24: 70–76, illus.
(227) PENNINGTON, M. E.
1932. FLAVOR AND EATING QUALITY. REFRIGERATED OR RECENTLY LAID EGGS —WHICH DO YOU PREFER? U. S. Egg and Poultry Mag. 38 (9): 28–31.
(227a) [PERET, H.]
1949. TASTE TEST PANELS. Natl. Provisioner 121 (20): 12–13, illus.
(228) PLANK, R. P.
1948. A RATIONAL METHOD FOR GRADING FOOD QUALITY. Food Technol. 2: 241–251, illus.
(229) PLATT, W.
1931. RATIONAL METHODS OF SCORING FOOD PRODUCTS. Food Indus. 3: 108–111, illus.
(230) PLOWMAN, [?]
1944. THE ART OF TEA-TASTING. Chem. and Indus., no. 29, p. 263.
(231) PORTER, T., SCHLAPHOFF, D. M., WHARTON, M. A., and others.
1944. ASCORBIC ACID CONTENT, COLOR, AND PALATABILITY OF FRESH AND PROCESSED SWISS CHARD AND BEET GREENS. Food Res. 9: 268–277.
(232) PUNNETT, P. W., and EDDY, W. H.
1930. WHAT FLAVOR MEASUREMENT REVEALS ABOUT KEEPING COFFEE FRESH. Food Indus. 2: 401–404, illus.
(233) RAMSBOTTOM, J. M., STRANDINE, E. J., JENSEN, L. B., and others.
1947. THE EFFECT OF 40 YEARS' FROZEN STORAGE ON THE QUALITY OF BEEF. Refrig. Engin. 54: 544–548, 583, illus.

(234) _____ STRANDINE, E. J., and KOONZ, C. H.
1945. COMPARATIVE TENDERNESS OF REPRESENTATIVE BEEF MUSCLES. Food Res. 10: 497–509, illus.
(235) REDGROVE, H. S.
1943. FLAVOUR EVALUATION. ESSENCE AND BACKGROUND. Food 12: 35–36.
(236) RICHTER, C. P., and CAMPBELL, K. H.
1940. SUCROSE TASTE THRESHOLDS OF RATS AND HUMANS. Amer. Jour. Physiol. 128: 291–297, illus.
(237) _____ and CLISBY, K. H.
1941. PHENYLTHIOCARBAMIDE TASTE THRESHOLDS OF RATS AND HUMAN BEINGS. Amer. Jour. Physiol. 134: 157–164, illus.
(238) RIDDELL, W. J. B., and WYBAR, K. C.
1944. TASTE OF THIOURACIL AND PHENYLTHIOCARBAMIDE. Nature 154: 669
(239) ROESSLER, E. B., WARREN, J., and GUYMON, J. F.
1948. SIGNIFICANCE IN TRIANGULAR TASTE TESTS. Food Res. 13: 503–505.
(240) SAIR, L., and COOK, W. H.
1938. EFFECT OF PRECOOLING AND RATE OF FREEZING ON THE QUALITY OF DRESSED POULTRY. Canad. Jour. Res. Sect. D, Zool. Sci. 16: 139–152, illus.
(241) SALMON, T. N., and BLAKESLEE, A. F.
1935. GENETICS OF SENSORY THRESHOLDS: VARIATIONS WITHIN SINGLE INDIVIDUALS IN TASTE SENSITIVITY FOR PTC. Natl. Acad. Sci. Proc. 21: 78–83, illus.
(242) SATORIUS, M. J., and CHILD, A. M.
1938. PROBLEMS IN MEAT RESEARCH. I. FOUR COMPARABLE CUTS FROM ONE ANIMAL. II. RELIABILITY OF JUDGES' SCORES. Food Res. 3: 627–635.
(243) SCHOLL, F. M., and MUNCH, J. C.
1937. TASTE TESTS. IV. RELATIVE BITTERNESS. (Abstract) Amer. Pharm. Assoc. Jour. 26: 127–129.
(244) SCHREIBER, M. L., VAIL, G. E., CONRAD, R. M., and PAYNE, L. F.
1947. THE EFFECT OF TISSUE FAT STABILITY ON DETERIORATION OF FROZEN POULTRY. Poultry Sci. 26: 14–19.
(245) SCHWAB, A. W., and DUTTON, H. J.
1948. THE FLAVOR PROBLEM OF SOYBEAN OIL. III. A FOUR-SAMPLE, GLASS LABORATORY DEODORIZER. Amer. Oil Chem. Soc. Jour. 25: 57–59, illus.
(246) SCOTT BLAIR, G. W., COPPEN, F. M. V., and DEARDEN, D. V.
1941. A PRELIMINARY STUDY OF THE EFFECTS OF VARYING PITCHING CONSISTENCY AND RATE OF SCALD ON THE PHYSICAL AND CHEMICAL PROPERTIES OF CHEDDAR CHEESE AND ON THE FIRMNESS OF THE CHEESE AS JUDGED BY CHEESE-MAKERS, BAKERS AND OTHERS. Jour. Dairy Res. 12: 170–177.
(247) SHARP, P. F.
1941. SOME OF THE QUESTIONS RAISED IN AN ADDRESS ON FACTORS INFLUENCING THE FLAVOR OF MILK. Milk Plant Monthly 30 (2): 31–34.
(248) _____ STEWART, G. F., and HUTTAR, J. C.
1936. EFFECT OF PACKING MATERIALS ON THE FLAVOR OF STORAGE EGGS. N. Y. (Cornell) Agr. Expt. Sta. Mem. 189, 26 pp., illus.
(249) SHELLENBERGER, J. A.
1939. VARIATION IN THE BAKING QUALITY OF WHEAT DURING STORAGE. Cereal Chem. 16: 676–682, illus.
(250) SHIMER, S. R., and PURINTON, H. J.
1948. THE ASCORBIC ACID AND CAROTENE CONTENT OF FRESH AND FROZEN NEW HAMPSHIRE BERRIES. N. H. Agr. Expt. Sta. Tech. Bul. 92, 19 pp.
(251) SHOWALTER, H. A.
1945. TASTE AND FLAVOURS. Food in Canada 5 (1): 9–11, illus.
(252) SHREWSBURY, C. L., HORNE, L. W., BRAUN, W. Q., and others.
1942. CHEMICAL, HISTOLOGICAL AND PALATABILITY CHANGES IN PORK DURING FREEZING AND STORAGE IN THE FROZEN STATE. Ind. Agr. Expt. Sta. Bul. 472, 36 pp., illus.
(253) SLOCUM, R. R., LEE, A. R., SWENSON, T. L., and others.
1933. A STUDY OF EGG FLAVOR IN STORED OIL-TREATED EGGS. U. S. Egg and Poultry Mag. 39 (4): 14–17, 47, illus.
(254) SMART, H. F., and BRUNSTETTER, B. C.
1937. SPINACH AND KALE IN FROZEN PACK. I. SCALDING TESTS. II. MICROBIOLOGICAL STUDIES. Food Res. 2: 151–163.
(255) SMITH, O., NASH, L. B., and DITTMAN, A. L.
1942. POTATO QUALITY. VI. RELATION OF TEMPERATURE AND OTHER FACTORS TO BLACKENING OF BOILED POTATOES. Amer. Potato Jour. 19: 229–254, illus.

(255a) SNEDECOR, G. W.
1946. STATISTICAL METHODS APPLIED TO AGRICULTURE AND BIOLOGY. Ed. 4, 485 pp., illus. Ames, Iowa.
(256) SNYDER, L. H.
1931. INHERITED TASTE DEFICIENCY. Science (n. s.) 74: 151–152.
(257) SPENCER, D. A.
1929. JUDGING COOKED MEAT. Amer. Soc. Anim. Prod. Proc. 1928: 119–121.
(258) STADTMAN, E. R., BARKER, H. A., MRAK, E. M., and MACKINNEY, G.
1946. STORAGE OF DRIED FRUIT. INFLUENCE OF MOISTURE AND SULFUR DIOXIDE ON DETERIORATION OF APRICOTS. Indus. and Engin. Chem., 38: 99–104, illus.
(259) STEFFEN, A. H., HOPKINS, E. W., KLINE, R. W., and WHETZELL, G. H.
1943. A CHEMICAL METHOD FOR SCORING DRIED WHOLE EGGS. U. S. Egg and Poultry Mag. 49: 308–310, 334–336, illus.
(260) STEINBERG, M. P., WINTER, J. D., and HUSTRULID, A.
1949. PALATABILITY OF BEEF STORED AT 0° F. AS AFFECTED BY MOISTURE LOSS AND OXYGEN AVAILABILITY. Food Technol. 3: 367–369.
(261) STEVENSON, F. J., and WHITEMAN, E. F.
1935. COOKING QUALITY OF CERTAIN POTATO VARIETIES AS INFLUENCED BY ENVIRONMENT. Amer. Potato Jour. 12: 41–47.
(262) STEWART, G. F., BEST, L. R., and LOWE, B.
1943. A STUDY OF SOME FACTORS AFFECTING THE STORAGE CHANGES IN SPRAY-DRIED EGG PRODUCTS. *In* Inst. Food Technol. Proc. pp. 77–89, illus.
(263) ——— HANSON, H. L., and LOWE, B.
1943. PALATABILITY STUDIES ON POULTRY: A COMPARISON OF THREE METHODS FOR HANDLING POULTRY PRIOR TO EVISCERATION. Food Res. 8: 202–211, illus.
(264) ——— HANSON, H. L., LOWE, B., and AUSTIN, J. J.
1945. EFFECTS OF AGING, FREEZING RATE, AND STORAGE PERIOD ON PALATABILITY OF BROILERS. Food Res. 10: 16–27, illus.
(265) ——— LOWE, B., and MORR, M.
1941. POST-MORTEM CHANGES IN NEW YORK DRESSED POULTRY AT 35° F. U. S. Egg and Poultry Mag. 47: 542–544, 571–572, illus.
(266) STILLMAN, J. T., WATTS, B. M., and MORGAN, A. F.
1944. PALATABILITY STUDIES ON HOME-DEHYDRATED VEGETABLES. Jour. Home Econ. 36: 28–34.
(267) STUART, L. S., GREWE, E., and DICKS, E. E.
1942. SOLUBILITY OF SPRAY-DRIED WHOLE EGG POWDER. U. S. Egg and Poultry Mag. 48: 498–503, 524–526, illus.
(268) SWARTZ, V. W.
1938. TWO FURTHER SIMPLE OBJECTIVE TESTS FOR JUDGING CAKE QUALITY. Cereal Chem. 15: 247–250.
(269) SWEETMAN, M. D.
1931. THE SCIENTIFIC STUDY OF THE PALATABILITY OF FOOD. Jour. Home Econ. 23: 161–172.
(270) ———
1936. FACTORS AFFECTING THE COOKING QUALITY OF POTATOES. Maine Agr. Expt. Sta. Bul. 383: 297–387, illus.
(271) SWENSON, T. L.
1939. A SUMMARY OF STUDIES ON THE OILING OF EGGS. U. S. Bur. Chem. and Soils MC-58, 27 pp. [Processed.]
(272) THIESSEN, E. J.
1947. THE CULINARY QUALITIES AND NUTRITIVE VALUES OF POTATOES GROWN UPON DRY AND IRRIGATED LAND. Wyo. Agr. Expt. Sta. Bul. 280, 31 pp., illus.
(273) THISTLE, M. W., PEARCE, J. A., and GIBBONS, N. E.
1943. DRIED WHOLE EGG POWDER. I. METHODS OF ASSESSING QUALITY. Canad. Jour. Res. Sect. D, Zool. Sci. 21: 1–7.
(274) ——— REID, M., and GIBBONS, N. E.
1943. DRIED WHOLE EGG POWDER. V. DEFINITION AND PROPERTIES OF LOW GRADE EGG POWDERS. Canad. Jour. Res. Sect. D, Zool. Sci. 21: 267–269, illus.
(275) TOMKINS, R. G.
1948. GRADING AND QUALITY OF VEGETABLES. Chem. and Indus., no. 7, p. 107.
(276) ———
1949. ORGANOLEPTIC TESTS IN THE FOOD INDUSTRY. II. THE USE OF A "TASTING PANEL" FOR ASSESSING THE CULINARY QUALITY OF DRIED VEGETABLES. Chem. and Indus., no. 11, pp. 167–168.

(277) _____ MAPSON, L. W., and WAGER, H. G.
1946. THE DRYING OF VEGETABLES. V. THE EFFECT OF SCALDING IN STARCH OR SULPHITE SOLUTIONS ON THE LOSS OF CAROTENE AND THE DETERIORATION IN CULINARY QUALITY OF DRIED CARROT DURING STORAGE. Soc. Chem. Indus. Jour., Trans. and Commun. 65: 384–385.
(278) TRELEASE, R. D., and KOONZ, C. H.
1945. QUALITY OF EVISCERATED POULTRY OBTAINED FROM DEFROSTED, DRESSED STOCK. Food Res. 10: 373–378.
(279) TROUT, G. M.
1946. TIME REQUIRED FOR MAKING FLAVOR JUDGMENTS OF MILK. Jour. Dairy Sci. 29: 415–419, illus.
(280) _____ ANDERSON, E. O., BABCOCK, C. J., and others.
1948. AN ANALYSIS OF THE RESULTS OF THE 1947 COLLEGIATE STUDENTS' INTERNATIONAL CONTEST IN JUDGING DAIRY PRODUCTS. Jour. Dairy Sci. 31: 823–829.
(281) _____ DOWNS, P. A., MACK, M. J., and others.
1942. THE EVALUATION OF FLAVOR DEFECTS OF BUTTER, CHEESE, MILK AND ICE CREAM AS DESIGNATED BY DAIRY PRODUCTS JUDGES. Jour. Dairy Sci. 25: 557–569.
(282) _____ DOWNS, P. A., MACK, M. J., and others.
1943. COMPARATIVE STANDARDIZATION OF BUTTER, CHEESE, MILK AND ICE CREAM FLAVOR SCORING. Jour. Dairy Sci. 26: 63–68, illus.
(283) _____ and SCHEID, M. V.
1943. THE INFLUENCE OF SEVERAL FACTORS UPON THE FLAVOR OF FROZEN SWEET CREAM. Jour. Dairy Sci. 26: 609–618, illus.
(284) _____ and SHARP, P. F.
1937. THE RELIABILITY OF FLAVOR JUDGMENTS, WITH SPECIAL REFERENCE TO THE OXIDIZED FLAVOR OF MILK. N. Y. (Cornell) Agr. Expt. Sta. Mem. 204, 60 pp., illus.
(285) TUCKER, R. E.
1948. ACCEPTABILITY AND ASCORBIC ACID CONTENT OF FROZEN RHODE ISLAND VEGETABLES. R. I. Agr. Expt. Sta. Bul. 302, 24 pp., illus.
(286) VAIL, G. E., and CONRAD, R. M.
1948. DETERMINATION OF PALATABILITY CHANGES OCCURRING IN FROZEN POULTRY. Food Res. 13: 347–357.
(287) VOLZ, F. E., GORTNER, W. A., and DELWICHE, C. V.
1949. THE EFFECT OF DESICCATION ON FROZEN VEGETABLES. Food Technol. 3: 307–313, illus.
(288) WEAVER, E.
1939. PHYSIOLOGICAL FACTORS AFFECTING MILK FLAVOR, WITH A CONSIDERATION OF THE VALIDITY OF FLAVOR SCORES. Okla. Agr. Expt. Sta. Tech. Bul. 6, 56 pp., illus.
(289) WECKEL, K. G.
1941. PUT YOUR TASTE BUDS TO WORK ON YOUR FLAVOR PROBLEMS. Internatl. Assoc. Milk Dealers Bul. 34: 136–140.
(290) WHITACRE, J.
1945. INFLUENCE OF SALT ON THE JUDGING OF COOKED VEGETABLES. Natl. Coop. Proj. Conservation of Nutritive Value of Foods, Prog. Notes 5, 4 pp. (Tex. Agr. Expt. Sta.) [Processed.]
(291) _____ FRAPS, G. S., YARNELL, S. H., and OBERG, A. G.
1944. EATING QUALITY AND SOME ASPECTS OF COMPOSITION OF TURNIP GREENS AT SUCCESSIVE STAGES OF GROWTH. Food Res. 9: 42–55.
(292) _____ HAWTHORN, L. R., and YARNELL, S. H.
1939. LENGTHENING THE STORAGE PERIOD OF CUCUMBERS. Tex. Agr. Expt. Sta. Bul. 576, 23 pp., illus.
(293) WHITE, W., DOWNS, P. A., MACK, M. J., and others.
1940. CORRELATION BETWEEN GRADES ON SCORES AND GRADES ON CRITICISMS IN THE JUDGING OF DAIRY PRODUCTS. Jour. Dairy Sci. 23: 1–12, illus.
(294) WHITE, W. H., WOODCOCK, A. H., and GIBBONS, N. E.
1944. SMOKED MEATS. III. EFFECT OF MATURATION PERIOD ON QUALITY. Canad. Jour. Res. Sect. F, Technol. 22: 107–118.
(295) WILLS, R. F.
1946. ORGANOLEPTIC AND HISTOLOGICAL CHANGES IN EVISCERATED FROZEN POULTRY STORED UNDER VARYING CONDITIONS OF TEMPERATURE AND TIME. 272 pp., illus. [Unpublished master's thesis. Copy on file Iowa State Col. Libr., Ames.]
(296) WILSON, R. V., and SLOSBERG, H. M.
1942. METHOD DEVELOPED FOR GRADING A DEHYDRATED FOOD. Food Indus. 14 (9): 56–58, illus.

(297) WOOD, E. A.
1946. MEASUREMENT OF TASTE SENSITIVITY. What's New in Home Econ. 11 (2): 174–176, 179, 181, 183, 184, illus.
(298) WRIGHT, R. C., CALDWELL, J. S., WHITEMAN, T. M., and CULPEPPER, C. W.
1945. THE EFFECT OF PREVIOUS STORAGE TEMPERATURES ON THE QUALITY OF DEHYDRATED POTATOES. Amer. Potato Jour. 22: 311–323.
(299) ——— PEACOCK, W. M., WHITEMAN, T. M., and WHITEMAN, E. F.
1936. THE COOKING QUALITY, PALATABILITY, AND CARBOHYDRATE COMPOSITION OF POTATOES AS INFLUENCED BY STORAGE TEMPERATURE. U. S. Dept. Agr. Tech. Bul. 507, 20 pp., illus.
(300) ZIEGLER, E. C., and SCHOFIELD, E. H.
1945. PROPERTIES OF FOODS. Food Manufacture 20: 55.

PARTICIPANTS IN CONFERENCE

ADAMS, GEORGIAN. Administrator — Nutrition, Office of Experiment Stations, U. S. Department of Agriculture, Washington, D. C.
AMES, RUTH. Wallerstein Laboratories, New York, N. Y.
BATCHELDER, ESTHER. Head, Food and Nutrition Division, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.
BENNETT, GRACE. Assistant Professor of Foods and Nutrition, School of Home Economics, Pennsylvania State College, State College, Pennsylvania.
BOGGS, MILDRED. Food Technologist, Western Regional Research Laboratory, U. S. Department of Agriculture, Albany, Calif.
BOYDEN, RUTH B. Assistant in Home Economics, Department of Home Economics, University of Kentucky, Lexington, Ky.
BRANT, A. W. Food Technologist, Bureau of Animal Industry, U. S. Department of Agriculture, Beltsville, Md.
BRASTOW, VERA. Food Specialist, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.
BRIANT, ALICE M. Associate Professor of Foods and Nutrition, College of Home Economics, New York State College of Home Economics, Cornell University, Ithaca, N. Y.
CALVIN, LYLE. Institute of Statistics, University of North Carolina, Raleigh, N. C.
CARTWRIGHT, L. C. Foster D. Snell, Inc., New York, N. Y.
COVER, SYLVIA. Professor, Department of Rural Home Research, Texas A and M College, College Station, Tex.
COX, GERTRUDE. Head, Institute of Statistics, University of North Carolina, Raleigh, N. C.
CROCKER, ERNEST C. Arthur D. Little, Inc., Cambridge, Mass.
CULPEPPER, C. W. Plant Physiologist, Bureau of Plant Industry, Soils, and Agricultural Engineering, U. S. Department of Agriculture, Beltsville, Md.
DAWSON, ELSIE H. Food Specialist, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.
DOCHTERMAN, ELSIE F. Statistician, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.
DREISBACH, MARGARET B. Home Economist, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.
EBBS, JANE. Nutrition Consultant, Office of the Quartermaster General, Washington, D. C.
FENTON, FAITH. Professor of Foods and Nutrition, Department of Home Economics, New York State College of Home Economics, Cornell University, Ithaca, N. Y.
FIFIELD, COLBURN C. Food Technologist, Bureau of Plant Industry, Soils, and Agricultural Engineering, U. S. Department of Agriculture, Beltsville, Md.
GADDIS, ADAM. Chemist, Bureau of Animal Industry, U. S. Department of Agriculture, Beltsville, Md.
GILPIN, GLADYS L. Food Specialist, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.
GODDARD, VERZ R. Nutrition Specialist, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.
GRANT, FRED M. Dairy Technologist, Bureau of Dairy Industry, U. S. Department of Agriculture, Washington, D. C.
GREENWOOD, MARY L. Associate Professor of Foods and Nutrition, Department of Home Economics, University of Connecticut, Storrs, Conn.
HANNING, FLORA M. Professor of Home Economics, Department of Home Economics, University of Wisconsin, Madison, Wis.
HARDENBURG, R. E. Plant Physiologist, Bureau of Plant Industry, Soils, and Agricultural Engineering, U. S. Department of Agriculture, Beltsville, Md.

HARRIS, BETSY. Food Specialist, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.

HEINZE, PETER H. Plant Physiologist, Bureau of Plant Industry, Soils, and Agricultural Engineering, U. S. Department of Agriculture, Beltsville, Md.

HENING, J. C. Assistant Professor of Chemistry, Department of Food Science and Technology, New York State Agricultural Experiment Station, Geneva, N. Y.

HILLS, CLAUDE H. Chemist, Eastern Regional Research Laboratory, U. S. Department of Agriculture, Philadelphia, Pa.

HOLLINGER, MARTHA E. Associate Professor of Nutrition, Department of Chemistry and Biochemistry, Louisiana State University, University Station, Baton Rouge, La.

HOOD, MAUDE P. Professor, Foods and Equipment, Department of Home Economics, University of Georgia, Athens, Ga.

HUSSEMAN, DOROTHY L. Professor of Home Economics, Department of Home Economics, University of Wisconsin, Madison, Wis.

JONES, I. D. Professor of Horticulture, University of North Carolina, Raleigh, N. C.

JORDAN, RUTH. Associate in Home Economics, Department of Home Economics, Agricultural Experiment Station, Purdue University, LaFayette, Ind.

JUSTIN, A. CHRISTINE. Home Economist, Office of Experiment Stations, U. S. Department of Agriculture, Washington, D. C.

KIRKPATRICK, MARY E. Food Specialist, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.

KOCH, E. J. Agriculturist, Bureau of Plant Industry, Soils, and Agricultural Engineering, U. S. Department of Agriculture, Beltsville, Md.

KRAMER, A. Associate Professor of Horticulture, Department of Horticulture, University of Maryland, College Park, Md.

LAMB, JESSIE C. Food Specialist, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.

LIEBERMAN, M. Plant Physiologist, Bureau of Plant Industry, Soils, and Agricultural Engineering, U. S. Department of Agriculture, Beltsville, Md.

LOWE, BELLE. Professor and Research Professor of Foods and Nutrition, Iowa State College, Ames, Iowa.

MCLAREN, BARBARA A. Assistant Professor and Assistant Home Economist, Department of Home Economics, Washington State College, Pullman, Wash.

MCNALLY, EDMUND H. Biologist, Bureau of Animal Industry, U. S. Department of Agriculture, Beltsville, Md.

MASON, DAVID D. Statistician, Bureau of Plant Industry, Soils, and Agricultural Engineering, U. S. Department of Agriculture, Beltsville, Md.

MAYFIELD, HELEN L. Assistant Home Economist, Department of Home Economics, Montana State College, Bozeman, Mont.

MEYER, BERNADINE H. Professor of Foods and Institution Management, University of Tennessee, Knoxville, Tenn.

MILLER, CORA. Department of Home Economics, Ohio University, Athens, Ohio.

MONTAGUE, ELAINE H. Food Specialist, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.

MORRISON, W. W. Marketing Specialist, Production and Marketing Administration, U. S. Department of Agriculture, Washington, D. C.

MOSER, HELEN. Food Technologist, Northern Regional Research Laboratory, U. S. Department of Agriculture, Peoria, Ill.

MOUNTJOY, BEATRICE. Food Specialist, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.

MURPHY, ELIZABETH F. Assistant Biologist, Department of Horticulture, University of Maine, Orono, Maine.

NOBLE, ISABEL T. Professor of Home Economics, Department of Home Economics, University of Minnesota, University Farm, St. Paul, Minn.

OVERMAN, ANDREA. Food Specialist, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.

PARKS, ALBERT. Statistician, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.

PAUL, PAULINE. Associate Professor of Foods and Nutrition, Department of Home Economics, Michigan State College, East Lansing, Mich.

PERSONIUS, CATHERINE. Head, Department of Foods and Nutrition, College of Home Economics, New York State College of Home Economics, Cornell University, Ithaca, N. Y.

PERYAM, DAVID R. Food Research Division, Quartermaster Food and Container Institute, Chicago, Ill.

PRUDENT, INEZ. Associate Professor of Home Economics, Department of Home Economics, Ohio State University, Columbus, Ohio.

PURINTON, HELEN J. Assistant Professor and Assistant Chemist, Department of Agricultural and Biological Chemistry, University of New Hampshire, Durham, N. H.

REUBELT, V. A. Biologist, Bureau of Plant Industry, Soils, and Agricultural Engineering, U. S. Department of Agriculture, Beltsville, Md.

REYNOLDS, HOWARD. Bacteriologist, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.

RIBACK, B. L. Schenley Company, Empire State Building, New York, N. Y.

SCHLOSSER, GEORGIA. Food Specialist, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.

SCHOPMEYER, GRACE E. Food Specialist, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.

SHEPHERD, JANE. Statistician, Division of Special Surveys, Bureau of Agricultural Economics, U. S. Department of Agriculture, Washington, D. C.

SJÖSTRÖM, LOREN B. Arthur D. Little, Inc., Cambridge, Mass.

STANLEY, LOUISE. Agricultural Research Administration, U. S. Department of Agriculture, Washington, D. C.

TOEPFER, EDWARD. Chemist, Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture, Beltsville, Md.

VAIL, GLADYS E. Head, Department of Foods and Nutrition, Department of Home Economics, Kansas State College, Manhattan, Kans.

VAN DUYNE, FRANCES O. Associate Professor of Foods, Department of Home Economics, University of Illinois, Urbana, Ill.

WEBB, LAURA MAE. Economist, Bureau of the Budget, Washington, D. C.

WEIHE, H. D. Chemist, Bureau of Dairy Industry, U. S. Department of Agriculture, Washington, D. C.

WEIR, EDITH. Food Technologist, Bureau of Animal Industry, U. S. Department of Agriculture, Beltsville, Md.

WOLGAMOT, IRENE H. Home Economist, Bureau of Human Nutrition and Home Economics. U. S. Department of Agriculture, Beltsville, Md.

☆U. S. GOVERNMENT PRINTING OFFICE: 1951-927678