Studies in Taste Panel Methodology

AMIHUD KRAMER
University of Maryland, College Park, Md.
ELIZABETH F. MURPHY
University of Maine, Orono, Maine
ALICE M. BRIANT
New York State College of Home Economics, Ithaca, N. Y.
MARIAN WANG
Pennsylvania State University, University Park, Pa.
MARY E. KIRKPATRICK
Human Nutrition Research Division, U. S. Department of Agriculture, Beltsville, Md.

Taste panels are indispensable for evaluating flavor and detecting off-flavor, but time-consuming and expensive. For a most efficient panel, a variables (scoring or ranking) method and a small number of trained panelists should be used to detect differences. Nine samples or more per sitting are effective, particularly for mild-flavored products. A reference sample is useful and should be included as both known and unknown. Analysis of variance is most informative, particularly to detect possible panelist × sample interactions. A rapid statistical rank sum analysis is presented, and a sample number procedure for comparing efficiency of taste-testing methods. These results should aid in substantially reducing costs of taste panels capable of providing statistically valid information.

THE FIELD OF SENSORY (taste panel) evaluation of foods and beverages is one where "empiricism is rampant, definitions and terminology need to be established, methodology needs to be screened, evaluated, and developed, and the results of such tests need to be interpreted in significant terms" (*9*). During investigations of flavor changes due to pesticides undertaken by a group of experiment stations in the Northeast and the U. S. Department of Agriculture, there was the opportunity to study and compare various methods with the purpose of arriving at some definite conclusions, based on factual data, as to the relative efficiency of procedures for specific taste-testing purposes.

Identify or Score

In selecting a given procedure for a particular taste-testing problem, the operator must first decide whether a sample is merely to be identified, or evaluated on some predetermined scale of values.

If the purpose of the test is selection of one sample over another, or a decision as to whether a sample is acceptable or not, actual values are of no particular interest, so that it has been assumed that an attributes test—one which merely requires selection or identification of a sample—is sufficient. However, in studies carried out under this project (*10*, *12*), it was demonstrated that a multiple comparison test, in which each sample was scored, was several times more efficient in detecting flavor differences than the triangle test (selection of the odd sample), which was used as representing an attributes procedure (Table I). Such a conclusion could be predicted by theoretical statistical considerations. Thus, for example, Bowker and Goode (*3*) show that 20 evaluations under a variables (scoring or ranking) procedure are as informative as 40 evaluations under an attributes procedure (identification or selection). The advantage of

Table I. Summary of Triangular (Attributes) and Multiple Comparison (Variables) Taste Test Results

	Products									
	Tomato Juice		Tomatoes		Potatoes		Snap Beans		Lima Beans	
Treatments	Tri-angle	Mult. comp.	Tri-angle	Mult. comp.	Tri-angle	Mult. comp.	Tri-angle	Mult. comp.	Tri-angle	Mult. comp.
Toxaphene 2%	10	0.3	10	0.4	16[a]	1.2[b]	10	0.4	15	0.1
Methoxychlor 2%	16[a]	0.8[b]	12	0.0	13	0.3	10	0.8	18[a]	0.0
Malathion 4%	11	0.4	14	0.7	13	0.2	11	0.6	15	0.5
Methoxychlor 5%, resin 1%, sulfoxide 2%	11	0.7[b]	12	0.6	12	1.3[b]	17[a]	0.4	13	0.4
Dilan 2%, lindane 2%	17[a]	−1.0[c]	10	0.2	10	−0.2[c]	13	0.3	13	−0.4[c]
Methoxychlor 2%, dilan 2%	15	−0.3	20[a]	0.8	11	0.9	14	0.6	15	0.3
Check		0.1		0.5		0.4		0.6		0.6
L.S.D.		0.6		0.6		0.5		0.3		0.7

[a] Significantly different from check.
[b] Significantly better than check.
[c] Significantly poorer than check.
Triangle values indicate number of correct selections out of 30, with at least 16 required for significance.
Multiple comparison values represent average scores with positive values indicating acceptable flavor, and negative values indicating off-flavor.

the variables system over attributes increases as the number of samples increase, so that results from 100 tests by the variables plan are equal in precision to 450 by attributes; and 200 tests by the variables plan are equal to 1500 by attributes.

Attributes results in themselves provide no information on the nature of the difference, when a significant difference is found. Thus a triangle test result may point to the fact that a panel has been able to select the odd sample, but provides no information on the extent or importance of the difference.

Of the three results in Table I that were found significant by both methods, only the effect of dilan plus lindane on tomato juice may be said to point to a definite off-flavor, whereas the significant effects of methoxychlor on tomato juice and toxaphene on potatoes are apparently due to an improvement in flavor, as indicated by the variables test. Similar results could probably have been obtained if, in addition to a mere selection of the odd sample, the judges participating in the triangle taste panel were required to note whether the odd sample were better or poorer than, or equal to the paired samples (or vice versa). Actually, such additional notations would change the original triangle attributes test to the simplest form of a variables test, where a three-point scale is used. If this needs to be done, it appears more logical to utilize a variables design from the start rather than to patch a less efficient attributes method to provide the complete answer.

Attributes methods appear to be most successful with products approaching homogeneity within lots. Thus the most useful results (Table I) from the triangle tests were on tomato juice, a fairly homogeneous material. In off-flavor studies, however, it is occasionally impossible, or at least impractical, to obtain samples identical in all respects except for the treatment under study. In such cases an attributes method such as the triangle test becomes entirely inadequate. What is needed is an experimental design (*5*) by which variability resulting from other factors—location, maturity, etc.—may be determined and isolated from the variability in flavor associated with the application of the controlled treatments.

Comparison of Efficiency of Procedures

Utilizing variables procedures, in which the analysis of variance may be used on the direct or the transformed data (*2, 6*), a means was needed for comparing different procedures in order to determine their relative efficiency directly. This was accomplished by the use of the sample number ratio (SNR), which is the number of samplings, or tastings required to achieve a statistically significant difference, divided by the number actually used. The SNR may be calculated as follows:

$$\mathrm{SNR} = \left(\frac{\mathrm{LSD}}{R}\right)^2$$

where LSD is the least significant difference at the 1% ($p = 0.01$) level, for treatment means as calculated by Duncan (*7*) and R is the range among treatment means. When multiplied by the number of tastings performed in the experiment (N) the sample number (SN) required to achieve a statistically significant difference is obtained.

Thus, for example, for the tomato juice data in Table I, the least significant difference is 0.6, and the range among the treatment means is 0.8 — (−1.0), or 1.8, and the number of tastings in the experiment is 140. Hence,

$$\mathrm{SNR} = \left(\frac{0.6}{1.8}\right)^2 = 0.111$$

and $$\mathrm{SN} = (0.111)\quad(140) = 16$$

indicating that a minimum of 16 tastings would have been required to demonstrate a statistically significant difference in flavor among the tomato juice treatments tested.

Selection of Panelists

In the selection of individuals to serve on taste panels, the purpose of the taste test is again to be considered first. If the purpose is to obtain a consumer reaction only, a trained panel is not needed, and perhaps should be avoided. However, for the purpose of inspection, or analysis of differences, it may be important to select panelists who have a superior ability to detect differences and/or who show good agreement with consumer evaluation.

Just one or two screenings for selecting panelists who appear to have superior ability in detecting flavor differences, are apparently insufficient. The data in Table II are summarized from studies previously reported (*17*) and show that on the average, after a first screening of 28 candidates, the 12 who performed best originally did not perform more efficiently than all the original 28 candidates. This may have been due to the fact that most of these 12 top ranking performers arrived at this position largely by chance, so that in another test a different group may have been so selected. A second screening of the candidates resulted in a more efficient group. Further screening and training would undoubtedly have resulted in a still more efficient panel.

Results of two panels were then studied —an experienced panel, which came through repeated screenings, and an unknown panel with little or no experience. Results on detecting flavor differences of 10 varieties of green beans (Table III) indicate that individual panelists vary considerably in their ability to detect differences, so that for purposes of difference detection, it is advisable to use fewer, well trained tasters, and to replicate sufficiently in order to achieve the desired precision, rather than to use little or no replication, and a larger number of judges. As the total number of replicate tastings increases, there is no particular improvement in the efficiency of the trained panel, but considerable improvement in the untrained panel (Table III, B). These conclusions are in good agreement with those of Murphy, Covell, and Dinsmore (*13*), who found that a sample number of 34 was sufficient to detect flavor differences among strawberry varities when eight trained panelists were used, but 254 tastings were required when panelists were selected at random.

Table II. Summary Results with 27 Panels, of Effect of Screening Candidates for Taste Panels

	Av. No. of Candidates	Av. SN[a]
Untested panel	28	95
First screening	12	97
Second screening	6	88

[a] Sample number required to demonstrate statistically significant differences.

Table III. Summary Results Comparing Panelists of Proved Ability with Untrained Panelists

A. Panelist	Sample Numbers Required for Significance: Trained	Untrained
A	15	82
B	24	88
C	40	190
D	70	500

B. No. of Tastings per Expt.	Average Sample Numbers: Trained panel	Untrained panel
40	42	215
80	40	135
160	48	78

Table IV. Correlation Coefficients between Responses of Taste Panelists and Consumer Quality Evaluations

A. Panelist	Correlation Coefficients (r): Proved ability	Untrained
A	0.64	0.33
B	0.66	0.75
C	0.86	0.65
D	0.57	0.30

B. No. of Tastings per Expt.	Av. of 8 Panelists
40	0.84
80	0.81
160	0.86

Table V. Effect of Number of Samples Tasted per Sitting on Efficiency as Measured by Sample Number Required for Significance

No. of Tastings per Sitting	Average Sample Number Required for Significance			
	Canned squash	Cooked potatoes	Canned applesauce	Canned peaches
1	117			
2	50	88		
3	14	150	33	29
4		108		
5		76		
6	13			
7				
9	8		40	26
18	9		22	

Table VI. Effect of Panel Training on Numbers of Samples of Applesauce Tasted per Sitting

No. of Tastings per Sitting	Sample Numbers Required for Significance	
	Untrained panel	Trained panel
3	33	8
9	57	7
18	366	7

Table VII. Effect of Order or Position of Samples on Efficiency of Taste Testing to Ascertain Differences

Treatment	Check in First Position		Check in Middle Position		Check in Last Position	
	Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2
Check	+1.02	+0.25	+0.29	+0.31	+0.13	+0.29
Off-flavor treatment 1	−0.38	−0.83	−0.11	−0.38	+0.11	−0.36
Off-flavor treatment 2	−0.18	−0.78	0.00	−0.59	−0.13	−0.40
Sample number	5	21	62	31	150	52

Check should equal 1.00; any negative value should indicate off-flavor.

Table VIII. Effect of Reference Sample on Efficiency of Taste Panels

No. of Tastings per Sitting	Sample Number Required for Significance					
	Squash		Applesauce		Peas	
	No. ref.	Ref.	No. ref.	Ref.	No. ref.	Ref.
1	117	50			25	8
3	13	14	33	9	10	7
6	14	12				
9	9	6	40	25	15	10
18	12	6	371	22	23	20

Table IX. Relative Efficiency of Taste Test Panels Employing Scoring and Ranking Techniques

No. of Tastings per Sitting	Sample Number Required for Significance			
	Potatoes		Tomatoes	
	Scoring	Ranking	Scoring	Ranking
2	164	189		
3	187	264		
4	135	156		
5	133	65		
7			10	29

That agreement with consumer evaluation is not necessarily related to proved ability to detect differences is demonstrated in Table IV, which shows rather wide variations in correlations with consumer quality evaluations among both proved and untried panelists. Furthermore, added replication did not necessarily improve the correlations for any individual panelists, or for the panel averages. The correlations were substantially improved, however, by increasing the number of panelists. It may be concluded, therefore, that it is advisable to increase numbers of panelists and reduce replications, if it is desired to obtain an indication of consumer preference.

Number of Samples per Sitting

There are probably more confusion and difference of opinion regarding the number of samples that should be presented to a panelist at one sitting than any other single point in taste panel methodology. Opinions vary from only one at a sitting, held by some psychologists (*1*), to no more than two to be treated as pairs (*4*), or three in triangle tests (*14*), to larger numbers for the sake of economy and opportunity for optimizing flavor memory (*12*).

The results obtained from several separate studies conducted under this regional project must lead to the conclusion that for obtaining significant differences, at least, a single sampling technique is extremely inefficient, with the optimum number to be handled at any one sitting depending on the nature of the substance being tested. If the substance is a bland product such as squash or potatoes, a considerable number can be handled at one sitting. As shown in Table V, efficiency per tasting of squash seemed to improve up to nine per sitting, and was not reduced appreciably even when increased to 18 per sitting. A similar trend was apparent with potatoes, where the efficiency of five samples per sitting was greater than that of fewer numbers per sitting.

There was some indication that with less bland products such as apple sauce and peaches, the most efficient number of samples per sitting might be more limited. Thus, when a larger untrained panel was used, three samples per sitting were superior to nine, and the use of 18 samples at one sitting seemed to lead to utter confusion (Table VI). When a trained panel was used, on the other hand, the number of samples per sitting from three to 18 seemed to have little effect on efficiency. With peaches also, there was little difference in efficiency when three or nine samples were used per sitting. This is very much in agreement with results of Tompkins and Pratt (*15*), who found little difference in ability to discriminate between three and seven samples of orange juice per sitting but found seven per sitting more efficient from the handling standpoint.

Position of Samples

Placing the check (untreated or reference) sample first appears to have a beneficial effect on the efficiency of detecting the off-flavor samples. In a case with peaches (Table VII), when the check sample was treated first, a sample number of five was sufficient to indicate a significant difference in flavor between the check and an off-flavor sample. This increased to 62 and 150, respectively, when the check sample was placed in positions 2 and 3. With another panel this position effect was in the same direction, though not as striking.

It is not practical to place a check treatment always in position 1 and have it remain as an unknown. However, the availability of an identified reference or check treatment may serve a similar purpose. The data in Table VIII on squash, applesauce, and peas indicate very substantial improvement in efficiency when a reference sample is available.

In the event of the presence of an identified reference sample, this same sample must also be included as an unknown. It has been shown time and

again in these studies that the reference sample submitted to panels as an unknown will invariably be downgraded slightly as compared to the known reference. On a five-point scale of +2, +1, 0, −1, and −2, where the reference sample was scored as +1.0, the same sample as an unknown scored between +0.8 and +0.6.

Proponents of the paired comparison school have suggested that the use of an identified reference sample actually reduces a multisample analysis to a series of paired comparisons.

Scoring vs. Ranking

Where it is difficult to provide absolute values, or where absolute values are not very meaningful, as in quantitative sensory evaluations, a ranking procedure may be more appropriate than scoring. Tompkins and Pratt (*15*) found that "under the peculiar conditions of the tests" which they conducted, the ranking technique was more precise than scoring of frozen orange juice. Similar results were obtained in these studies with potatoes (Table IX), where the ranking technique appeared to be better when the number of samples per sitting was increased to five. In another study with canned tomato juice, however, results with scoring were definitely more efficient than those obtained by ranking. In general, therefore, these results cannot be interpreted to indicate a definite advantage favoring the use of either scoring or ranking procedures.

Preparation of Samples

Obviously samples should be prepared with care before presentation to the panelists. In general, from a statistical standpoint it is desirable to comminute materials which are not homogeneous, so that each panelist may be provided with an essentially similar aliquot. This generalization must be modified by the apparent fact that the product should be tasted by the panelist in the condition in which it is normally consumed. Thus when canned peaches were presented as sliced, diced, and puréed (Table X), the ability of the panel to detect flavor differences correctly was greatest when tasted as slices, almost as good when diced, but decidely poorer when puréed. It may be assumed that had the panel consisted of babies accustomed to eating puréed peaches, their efforts would have been most effective on the puréed samples.

It appears to be most important to mask any color or other differences not related to flavor, so that they may not be superimposed on flavor differences. This was brought out very strikingly when these same peaches were presented as frozen samples. Here results with the puréed samples were by far more precise; however, the differences had to do not with flavor but with color, since the pesticidal treatment which caused a flavor change also coincidentally affected the rate of browning of the frozen puréed peaches. Thus, the puréed samples which should have been scored as off-flavor were actually scored in some instances as superior because of peach color.

Table X. Relative Efficiency of Taste Panels Tasting Canned or Frozen Peaches

No. of Tastings per Sitting	Sample Number Required for Significance					
	Canned			Frozen		
	Sliced	Diced	Puréed	Sliced	Diced	Puréed
3	25	32	292	37	146	121
9	21	30	72	146	15	15

Statistical Analysis

Scoring—that is, variables—procedures lend themselves admirably to an analysis of variance (*12*), although there are certain objections to the use of this statistical procedure, since scores obtained from taste panels may not satisfy certain assumptions underlying the analysis of variance (*8*). When imposed upon a good experimental design, the analysis of variance will provide the opportunity for isolating and removing from the treatment effects all kinds of incidental sources of variations, including effects of time and position. This kind of procedure also provides for determining the importance of interactions. One extremely important interaction in taste panel results is the treatment × panelist interaction, which, when found to be significant, indicates that individual panelists score the same samples differently. This means in effect that there is no best sample or worst sample, but that a certain sample is considered better by some panelists, while another sample is preferred by others.

Particularly where the experimental design is not complicated, a number of rapid approximations of the variance method have been suggested, such as the range procedure (*16*).

Where ranking is to be used, ranks may be converted according to Fisher and Yates (*2*), and the usual analysis of variance carried out. Another rapid procedure in which the ranks may be used directly was developed as part of this project. This procedure, based on the multinomial distribution, is presented elsewhere (*11*).

Summary

For use with taste panels required to detect flavor differences, a variables method—that is, a scoring or ranking procedure—was more efficient than an attributes procedure, such as the triangle test, which requires the mere selection or identification of a sample.

Expert panelists cannot be obtained after one screening but need training and experience in repeated screening before they achieve expertness in detecting differences. A small number of trained panelists is adequate for detecting flavor differences, and the degree of precision required may be achieved by replicated tastings of the same samples. For determination of consumer preference, however, such expert panelists are not needed, and precision may be attained only by increasing the number of panelists, rather than replication.

For detecting differences, presentation of one sample per sitting is highly inefficient. Considering total effort expended, for mild-flavored products at least, there appears to be little reason for not using as many as nine samples per sitting. With trained panels the number of samples per sitting can probably be even higher, with a net gain in efficiency.

Where an identified reference sample is not available, placing the reference sample in first position improves the chances of detecting differences. An identified reference sample, where one is available, is very beneficial, but where used, the same sample should also be included as an unknown.

Neither scoring nor ranking was found definitely superior. Careful preparation of samples is of real importance, with particular attention to the possible confounding of visual characteristics with flavor differences. Scores obtained from experiments designed statistically and analyzed by the analysis of variance give maximum information, including some highly important interactions such as the one between panelists and samples. Particularly with simple, standard designs, some quick statistical procedures are available. A rank sum procedure has been developed, particularly where ranking is used instead of scoring.

The sample number (SN) is proposed as a means of comparing precision of different taste-testing procedures.

Literature Cited

(1) Bayton, J. A., Trans. Mid-Atlantic Conference, Am. Soc. Quality Control, 50 Church St., New York, N. Y., pp. 23–8, 1956.
(2) "Biometrika Tables for Statisticians," 2nd ed., Vol. **1**, p. 175, Cambridge University Press, London, 1958.

(3) Bowker, A. H., Goode, H. P., "Sampling Inspection by Variables," McGraw-Hill, New York, 1952.
(4) Byer, A. J., Abrams, Dorothy, *Food Technol.* **7,** 185–7 (1953).
(5) Cochran, W. G., Cox, G. M., "Experimental Designs," Wiley, New York, 1957.
(6) Dawson, E. H., Harris, B. H., U. S. Dept. Agr. Bull. **34,** 108–12 (1951).
(7) Duncan, D. B., *Biometrics* **11,** 1–42 (1955).
(8) Eisenhart, Churchill, *Ibid.*, **3,** 1–21 (1947).
(9) *Food Technol.* **13,** 733–6 (1959).
(10) Hogue, D. V., Briant, A. M., *Food Research* **22,** 351–7 (1957).
(11) Kramer, A., *Food Technol.* **14,** 576–81 (1960).
(12) Kramer, A., Ditman, L. P., *Ibid.*, **10,** 155–9 (1956).
(13) Murphy, E. F., Covell, M. R., Dinsmore, J. S., Jr., *Food Research* **22,** 423–9 (1957).
(14) Peryan, D. R., *Ind. Quality Control* **6,** 11 (1950).
(15) Tompkins, M. D., Pratt, G. B., *Food Technol.* **13,** 149–52 (1959).
(16) Tukey, J. W., *Proc. Am. Soc. Quality Control* **5,** 189–97 (1951).
(17) Wiley, R. C., Briant, A. M., Fagerson, I. S., Sabry, J. H., Murphy, E. F., *Food Research* **22,** 192–205 (1957).

Received for review August 18, 1960. Accepted December 12, 1960. Scientific article A845, Contribution 3136, Maryland Agricultural Experiment Station, Department of Horticulture.

PESTICIDES AND FOOD FLAVO

Influence of Herbicides on Flavor of Processed Fruits and Vegetables

F. J. McARDLE and A. N. MARETZKI

Pennsylvania Agricultural Experiment Station, University Park, Pa.

R. C. WILEY and M. G. MODREY

Maryland Agricultural Experiment Station, College Park, Md.

As a part of regional research on flavor effects of pesticides, 28 herbicides were applied to major processing crops. Manufacturers' suggested rates were used with all chemicals, and some were applied in excess of the suggested rate to increase effectiveness of weed control. Flavor evaluation of processed products by experienced taste panels indicated that 11 herbicides reduced product flavor scores; two produced slight off-flavors when applied at their suggested rates; three produced slight off-flavors when applied in excess; 17 of the chemicals studied did not reduce flavor scores of any products treated. The flavor changes observed were of low magnitude and might not have been detected by a consumer panel.

FLAVOR CHANGES in processed fruits and vegetables caused by the use of pesticides (insecticides, fungicides, and herbicides) on growing crops have been noted frequently during the past decade (*1–3*). These may result from application of the pesticide to the crop during its growing season or from an accumulation in the soil of pesticide residues from past seasons.

A study of the influence of pesticides on the flavor of fresh and processed fruits and vegetables was initiated on a regional basis in 1954 at the agricultural experiment stations in the northeast region. The effects of herbicides on the flavor of processed fruits and vegetables were studied cooperatively at The Pennsylvania State University and The University of Maryland.

Procedure

Herbicides were applied to major processing crops grown on horticultural farms at the Pennsylvania and Maryland stations. Manufacturers' suggested rates and methods of application were followed when this information was furnished for the chemical. Rates for

Table I. Taste Panel Flavor Evaluation of Herbicide-Treated Crops

Herbicide	Food Product	Application Rate, Lb./Acre	Crop Years	Flavor Compared to Standard: Better	Equal	Poorer	Slight off-flavor
				Number of Tests			
ACP 103	Corn, canned	1.5	2		3	1	
	Total				3	1	
ACP M 118	Lima beans, canned	3.0	1	1	3		
		4.5	2		6		
	Total			1	9		
ACP M 119	Lima beans, canned	2.0	1	1	1		
		3.0	2	1	5		
	Total			2	6		
ACP M 622	Tomatoes, canned	10.5	1		2		
	Total				2		
Atrazine	Corn, canned	2.0	1		4		
	Total				4		
Benzac 103 A	Corn, canned	1.5	1		1	1	
	Total				1	1	
Chlorazine	Lima beans, canned	4.0	1	1	1		
		6.0	2	1	5	2	
	Corn, canned	12.0	1		1	1	
	Total			2	7	3	
CIPC (Chloro-IPC)	Beets, canned	3.0	2		4		
	Carrots, canned	6.0	1	1	1		
	Lima beans, canned	4.0	1	1	1		
		6.0	2	1	5		
	Spinach, frozen	3.0	1		2		
		6.0	1		2		
	Tomatoes, canned	6.5	1		2		
	Total			3	17		