University of Connecticut
OpenCommons@UConn

Storrs Agricultural Experiment Station

College of Agriculture, Health and Natural Resources

11-1943

Technique for Testing Consumer Preferences, with Special Reference to the Constituents of Ice Cream, A

C.I. Bliss
University of Connecticut - Storrs

E.O. Anderson
University of Connecticut - Storrs

R.E. Marland
University of Connecticut - Storrs

Follow this and additional works at: https://opencommons.uconn.edu/saes

Part of the Food Chemistry Commons, Food Processing Commons, and the Other Food Science Commons

Recommended Citation
Bliss, C.I.; Anderson, E.O.; and Marland, R.E., "Technique for Testing Consumer Preferences, with Special Reference to the Constituents of Ice Cream, A" (1943). *Storrs Agricultural Experiment Station*. 34.
https://opencommons.uconn.edu/saes/34

Bulletin 251 November, 1943

STORRS
Agricultural Experiment Station

A TECHNIQUE FOR TESTING CONSUMER PREFERENCES, WITH SPECIAL REFERENCE TO THE CONSTITUENTS OF ICE CREAM

C. I. Bliss, E. O. Anderson and R. E. Marland

UNIVERSITY OF CONNECTICUT
STORRS, CONNECTICUT

FOREWORD

In the manufacture of ice cream, the selection of amounts and kinds of ingredients is an ever present problem. Aside from the nutritive values, the final test is the reaction of the consumer. Originally, the experiments herein reported sought merely to determine what our students prefer in ice cream. Then arose the need to determine the significance of the differences, if any, and to express the results objectively. However it soon became apparent that from such an experiment, properly designed, might be devised a better technique for testing consumer preferences. The aid of Dr. Bliss was sought and the statistical technique is his. It is presented in sufficient detail to be followed by the average experimenter.

Tests for consumer preferences are often needed for both natural and manufactured foods, for textiles and other products. The plant breeder, the horticulturist, the home economist, as well as those interested in ice cream and other dairy products, should find this bulletin of interest. In fact, because of their wide usefulness, the design of choice tests and their statistical analysis became a dominant objective.

W. L. Slate.

A TECHNIQUE FOR TESTING CONSUMER PREFERENCES, WITH SPECIAL REFERENCE TO THE CONSTITUENTS OF ICE CREAM.

C. I. Bliss[1], E. O. Anderson[2] and R. E. Marland[3]

In the standardization of food products, consumer preferences are often an important consideration. Most tests of palatability have been made by trained observers. In judging dairy products, for example, samples have been scored for each of many, named flavor defects, which in the case of butter fall in 25 different categories (12). Individuals trained to recognize off-flavors rated a series of products consistently (11). A simpler scoring system with three principal criteria has been described for judging the culinary quality of white potatoes (2) and in the hands of experienced workers gave concordant results.

How well these ratings would agree with ranks assigned by untrained observers may be questioned. Instead of scoring the different samples separately for each of several qualities which are later combined, frequently without weighting, it is simpler to grade them for a single dominant characteristic or without identifying the factors which determine preference. This technique may be used either with trained observers or with laymen. White *et al.* (13) have shown by correlation studies that student judges who were unable to criticize dairy products accurately might still score them reliably, suggesting that valid judgments of rank need not depend upon the ability to explain a preference in words.

Judgments of quality which are based on a fixed grading system, however, may show a high degree of observer's bias. This was demonstrated in tests reported by Stevenson and Whitman (10), where the quality of certain potato varieties grown in different locations was scored on a scale of 1.0 for very poor, to 5.0 for very good. Analysis of variance showed a significant difference between the mean scores of the five observers but consistency in following their own standards. In tests on the palatability of sweet corn, Dove (4) has avoided this observer's bias by a ranking technique. In any given test the observer received a one-inch section from each of six different varieties for ranking in order of choice. Palatability was clearly a relative term and the rating of a given variety depended upon the other varieties in the test.

The present experiments used an experimental technique similar to that described by Dove. They concern consumer reactions to certain components in ice cream, but the statistical techniques, described

[1] Consulting biometrician.
[2] Associate Professor, Department of Dairy Industry.
[3] Former 4-year student, Department of Dairy Industry.

in detail, are applicable in other fields where the subjects rank a given series of objects in order of choice. The procedure is adapted either to untrained groups, representative of the "average" consumer, or to experienced judges. Even where distinct grading systems have been developed, the use of ranks may permit discrimination between items which ordinarily would be grouped into a single class. Suitable tests of significance enable the experimenter to test the reliability of the conclusions reached by these numerical, objective techniques. Although each of the present experiments involved only four items, the method is not limited to series of four but may be extended to as many as can be handled effectively by the experimenter without fatiguing the subject.

Experimental data. The experiments tested four ingredients of ice cream, each in an independent series with all other constituents constant. In the two tests on flavorings, the ice cream contained 11 percent serum solids and 14 percent of fat. In the test on the concentration of serum solids the ice cream contained 14 percent of fat and in that on fat content it contained 11 percent of serum solids, both series being flavored with the same amount of pure extract of vanilla. All mixes contained 15 percent of cane sugar and 0.3 percent of gelatin (Swifts viscomix). They were pasteurized at 180° F. for 19 seconds by the Electropure shorttime-hightemperature pasteurizer, homogenized at 2,500 pounds pressure, cooled to 50° F., aged 48 hours and frozen to 90 percent overrun in a batch freezer. Samples of ice creams were collected in five-gallon ice cream cans and the choice tests were made within 48 hours after freezing.

For each variable four alternatives were prepared. The test on chocolate flavor compared American process flavoring of fountain quality with "Olympia," "Velvetier" and "Carbo" grades of Dutch process chocolate. That on vanilla flavoring contrasted natural vanilla, artificial vanilla, a 50-50 mixture of both at the same total concentration, and no flavoring. The experiment on serum solids compared concentrations of 8, 10, 12 and 14 percent, and that on fat content, concentrations of 8, 10, 14 and 18 percent.

For each series the four alternative ice creams were presented to the subject in small, lettered containers. He was instructed to taste the ice cream in all cups in any order he desired and with such repetition as was necessary to arrange them in order of choice. The subjects were students at the University of Connecticut totalling 58 boys and 12 girls, half of them participating in three or more tests. Although many of the group were agricultural majors, they represented both rural and urban backgrounds. The social, economic and regional characteristics of the subjects, of course, largely determine the inferences which may be drawn from the experiment.

Although the use of trained observers would give a biassed picture of the consumer's choice, an experiment will be more reliable if the subjects react consistently. This can be determined by testing each participant with the same alternatives, lettered differently, on two separate occasions. By computing the correlation coefficient between the scores reported in the two tests, individuals who have no consistent preferences for the same series of food products on two different occasions can be identified. Such data might well be segregated in the analysis of an experiment.

The value of this criterion was not appreciated until after the present experiments were completed, so that duplicate tests are available on only 6 girls for the series with a variable fat content. After transformation to scores as discussed below, these showed correlation coefficients of $r = -0.54$, 0.23, 0.77, 0.84, 1.00 and 1.00 respectively. With only two degrees of freedom r has a large error, yet it is evident that 4 of these 6 students were suitable subjects for an experiment on consumer preferences, while to the two girls with coefficients of -0.54 and 0.23 the different ice creams either tasted alike or varied in their desirability from one day to the next. Since there were so few duplicate tests, the analysis included the scores of all subjects without selection.

The transformation of ranks to scores. The advantages of the analysis of variance are well known and need not be discussed here. However, the distribution of simple ranks such as 1, 2, 3 and 4 departs more from the normal form than one would prefer for direct use in the analysis of variance. First and last choices, for example, tend to be ranked more easily than the intermediate items in a series. Fisher and Yates in Table XX(5) have provided a normalizing transformation for ordinal or ranked data which corrects this tendency. In a series of given size, each item is assigned a score equal to the expected value for an observation of corresponding rank in a normal population with a mean of zero and a standard deviation of one. Scores have been prepared for series of all sizes from 2 to 50 items. Since these are measured symmetrically from a mid-point of zero, the total score for each subject is zero with one less degree of freedom than the number of items in the set. In the case of ties, the corresponding scores are averaged, but in the absence of ties, a supplementary table (No. XXI in ref. 5) gives the sum of squares for the scores of each subject. The transformed scores are suitable for both the analysis of variance and the estimation of the correlation coefficient as reported in the preceding paragraph. For a different approach, which applies χ^2 to ranked data, the reader is referred to papers by Friedman (6) and by Kendall and Smith (8).

In the present experiment every series consisted of four alternatives and the subjects were not permitted to report ties. From reference (5) the four choices 1, 2, 3 and 4 were scored as 1.03, 0.30, -0.30 and -1.03 respectively, the sum of squares for each subject being 2.3018 with three degrees of freedom. The four series of tests were then examined by the analysis of variance to determine whether consumer preferences existed within this group of students concerning each characteristic under study, to isolate sex differences if present and finally to learn the preferred component or concentration in each series. The analysis of the two qualitative series on flavorings differs somewhat from that of the percentage concentrations of fat and serum solids, which will be considered separately.

The comparison of qualitative differences. The basic analysis may be described in detail for the series on four chocolate flavorings. The original ratings of four preparations for 32 boys and 11 girls are summarized in the frequency distributions of Table 1.

TABLE 1.

Frequency distributions of scores in tests of chocolate flavorings

Flavor	Boys					Girls					Total	
	1.03	.30	–.30	–1.03	*S(fy)*	1.03	.30	–.30	–1.03	*S(fy)*	*S(fy)*	$[y^2]$
American process	18	9	0	5	16.09	5	1	2	3	1.76	17.85 = A	25.5960
Dutch Olympia	4	6	8	14	–10.90	1	0	4	6	–6.35	-17.25 = B	20.7640
Dutch Velvetier	3	4	20	5	–6.86	5	1	4	1	3.22	–3.64 = C	15.0494
Dutch Carbo	7	13	4	8	1.67	0	9	1	1	1.37	3.04 = D	19.1466

Designating each transformed score by y and the number of "votes" for any given score or the frequency by f, the total score, $S(fy)$ may be computed for the boys and girls separately and then totalled for each flavoring. The sums for the flavorings have been designated by the letters A, B, C, and D respectively and must total zero.

The variability represented in the frequency distributions of the table is to be subdivided into three main portions: (1) that due to differences between the four flavorings, (2) that representing a sex difference in choice and (3) the residual variation or error. The design of the experiment divides the first of these into two sections, the difference between American and Dutch process and that between the three qualities or brands of Dutch process chocolate. The variance for the first comparison was computed from the difference 3A-B-C-D as

$$\frac{(3A\text{–}B\text{–}C\text{–}D)^2}{12N} \quad . \ . \ . \ . \ . \ . \quad (1)$$

where 12 is the sum of the squares of the coefficients in the numerator $(3^2+1+1+1)$ and $N=43$ or the total number of subjects. Numerically, we find

$$[(3 \times 17.85) + 17.25 + 3.64 - 3.04]^2/516 = 9.8798,$$

which is the first entry in the analysis of variance in Table 2.

TABLE 2.

Analysis of variance for the data in Table 1

	Degrees of freedom	Sum of squares	Mean square	*F*
American vs. Dutch process	1	9.8798	9.8798	15.09
Between Dutch chocolates	2	4.9731	2.4866	3.80
Boys vs. girls	3	3.5684	1.1895	1.82
Error	123	80.5561	.6549	
Total	129	98.9774		

This first treatment effect may then be subtracted from the total sum of squares for treatments or

$$\frac{A^2+B^2+C^2+D^2}{N} \quad . \; . \; . \; . \; . \; . \quad (2)$$

to obtain the sum of squares between the Dutch chocolates. Note that since the scores total zero, the usual correction for the mean is unnecessary. Numerically, we have

$$(17.85^2 + 17.25^2 + 3.64^2 + 3.04^2) \;/\; 43 - 9.8798 = 4.9731,$$

the second entry in Table 2.

The highest scoring chocolate among the boys was American process (16.09) but among the girls it was Velvetier Dutch process flavoring (3.22). Was the sex difference between flavorings large enough to be considered significant statistically in view of the numbers involved? The sum of squares for this comparison may be computed readily by applying Equation (2) separately to the totals for boys, dividing by the number of male subjects, and to the totals for girls, dividing by the number of girls. The sum of squares for boys plus that for girls is then diminished by the sum of squares computed from the total scores for both sexes. The difference is the sum of squares for the discrepancy between sexes with three degrees of freedom. Numerically the contrast between boys and girls may be computed as

$$\frac{(16.09^2+10.90^2+6.86^2+1.67^2)}{32} + \frac{(1.76^2+6.35^2+3.22^2+1.37^2)}{11} - 9.8798$$

$-4.9731 = 3.5684$, for the third row of Table 2.

The total sum of squares is equal to the product of the sum of squares for a single subject (2.3018) multiplied by the number of subjects (43) or $2.3018 \times 43 = 98.9774$ with $3 \times 43 = 129$ degrees of freedom. Subtracting the first three items to obtain the error, we find the mean square or variance for error is equal to $80.5561/123 = 0.6549$. From the ratio (F) of the first three mean squares to the error, we find from appropriate tables, such as from Table V for the variance ratio in reference (5) or from Table 10.3 in reference (9), that the preference for the American over Dutch chocolates was highly significant ($P < .001$), the difference between the three Dutch process brands significant ($P = .03$) but differences in the preferences of the boys and girls were not large enough to be considered established ($P = .16$).

The original data for the comparison of natural and artificial vanilla flavorings are summarized in Table 3 and analyzed in Table 4.

TABLE 3.

Frequency distributions of scores in tests of natural and artificial vanilla flavoring

	Boys					Girls					Total	
Flavor	1.03	.30	–.30	–1.03	$S(fy)$	1.03	.30	–.30	–1.03	$S(fy)$	$S(fy)$	$[y^2]$
Both types	15	7	2	5	11.80	6	0	1	2	3.82	15.62 = A	24.1824
Natural	6	9	10	4	1.76	2	4	2	1	1.63	3.39 = B	15.6397
Artificial	3	6	4	16	–12.79	0	3	5	1	–1.63	–14.42 = C	16.9020
No flavoring	5	7	13	4	– .77	1	2	1	5	–3.82	–4.59 = D	16.3417

TABLE 4.

Analysis of variance for the data in Table 3

	Degrees of freedom	Sum of squares	Mean square	F
Between flavorings	3	12.7495	4.2498	6.28
Boys vs. girls	3	1.6532	.5511	.81
Error	108	73.0657	.6765	
Total	114	87.4684		

The calculations differ from those for the chocolate ice creams only in the subdivision of the effects of treatment. Instead of comparing four different flavors, the experiment tested the preferences for natural and artificial flavoring, alone and in a 50-50 mixture, and no flavoring at all. The series differs from the customary 2 × 2 factorial design in that the ice cream with both flavorings contained only a half dose of each, a necessary modification to avoid confounding a qualitative factor with the effect of the amount of flavoring. The test is incomplete in that the artificial and natural flavors were not tested initially or in the same test at several concentrations. As shown in Table 4, subjects discriminated significantly between the four differently flavored ice creams, there being less than one chance in 1000 that a difference as marked as this could occur fortuitously. Both sexes reacted substantially alike, the variability between them being less than the error.

Since the discrimination between flavorings was so well established, it is useful to determine how large the difference between any two total scores would need to be before it could be considered significant. This may be computed as

significant difference $= t\sqrt{2Ns^2}$, ... (3)

where N is the number of observers, s^2 is the mean square for error and t is the ratio of a difference to its estimated error (with n degrees of freedom) for the desired level of significance. Tables of the statistic t are given by several authors including references (5 and 9), and it is customary to use the level for $P = .05$ in computing the just significant differences. For 108 degrees of freedom at $P = .05$, $t = 1.984$, so that here the significant difference $= 1.984\sqrt{2 \times 29 \times .6765} = 14.23$ between total scores. When used alone, the natural vanilla at these concentrations was preferred to the artificial flavor, but the mixture of the artificial and natural flavors was preferred to either alone although not significantly more than the natural vanilla. The ice cream without flavoring differed significantly only from the mixture of both types.

The analysis of quantitative factors. Two of the experiments tested consumer preference for differing percentages of fats and serum solids. The subjects discriminated significantly between the four levels of constituent in both series. In each case the percentage of constituent receiving the highest score was a mid-value, the scores for lower and higher percentages decreasing the more they departed from that with the highest score. The three total scores which bracketed the preferred concentration represent equally spaced percentages. Given these and certain other conditions, it is easy to compute the preferred percentage of constituent and its standard error. If the scores were equal for the two preparations on either side of that with the largest score, the maximum would coincide with the percentage of constituent in the sample tested experimentally. Since the outside scores differed from each other, the preferred percentage must be interpolated from a curve fitted to three or more points.

The curve computed from the observed scores is necessarily an empirical one and the equation best suited for this purpose is the parabola. It can be computed efficiently by least squares and if the successive concentrations of the ingredient are spaced equally on an arithmetic or logarithmic scale, the calculation can be simplified with orthogonal coefficients. The equation has the general form

$$y = a + b_1x + b_2x^2 \qquad \ldots \ldots (4)$$

and the value of x, measured from $\bar{x}$, for which y is a maximum (x_m) is given by differentiation as

$$x_m = -\frac{b_1}{2b_2} \qquad \ldots \ldots (5)$$

The sampling errors resulting from differences in individual preference diminish as more subjects are tested, but increase if the concentrations of the ingredient are chosen too near to x_m. However, as the range of concentrations is increased, the discrepancy increases between the curve defined by Equation (4) and the true but unknown relation of y to x. Hotelling (7) discusses mathematically the selection of the most efficient intervals of x for determining the maximum.

Assuming that they fall above and below the maximum, three concentrations are the minimum number which can be used. The parabola then passes through the three mean scores. However, with only three concentrations we are unable to determine whether the fitted curve agrees with the observations as closely as would be expected from the variation of the individual scores about their respective means, just as we cannot test the adequacy of a straight line when it is fitted to only two points. By fitting Equation (4) to the scores at four or more concentrations, one or more degrees of freedom are available for testing the suitability of the parabola. If the observations differ significantly from the fitted curve, it may be preferable in confirmatory experiments to restrict the range enclosing the maximum. This tends to minimize the discrepancy between the curve defined by Equation (4) and that expressing the true relation between x and y. Four or five concentrations, spaced symmetrically at equal intervals about the expected maximum, would be the preferred distribution of treatments in experiments of the present type for estimating the value of x having the highest preference.

The calculation will be described in detail only for evenly spaced concentrations. Let A, B, C, D, E stand for the sums of the scores for successive concentrations of ingredient in an ascending order, I

the interval in percentages between the equally spaced concentrations, $\bar{x}$ the mean percentage of ingredient over all values used in computing the maximum, N the number of individuals participating in the experiment and s^2 the pooled variance for all preparations in units of the individual score. We will assume that the variance in the response is the same for all concentrations of the ingredient under study, an assumption that will be examined in the next section.

The first step is to test the significance of the parabola fitted to the total scores by computing the variance accounted for by b_1 and by b_2 in Equation (4). When the concentrations are evenly spaced on an arithmetic or logarithmic scale, these variances are independent of one another and may be designated by $[L^2]$ for the linear term and by $[Q^2]$ for the quadratic term, each with one degree of freedom. Their equations may be expressed in tabular form as

Symbol	For 3 concs.	For 4 concs.	For 5 concs.	Equation No.
$[L^2]$	$\frac{(C-A)^2}{2N}$	$\frac{(3D+C-B-3A)^2}{20N}$	$\frac{(2E+D-B-2A)^2}{10N}$	(6)
$[Q^2]$	$\frac{(2B-A-C)^2}{6N}$	$\frac{(B+C-A-D)^2}{4N}$	$\frac{(B+2C+D-2A-2E)^2}{14N}$	(7)

The ratio of $[Q^2]$ to s^2, the error for the experiment as a whole, should exceed the value of F at $P = .05$, as given by standard tables (5,9) for $n_1 = 1$ and $n_2 =$ degrees of freedom in s^2. If not clearly significant, the range of concentrations may have been too short or misplaced above or below the optimal value. Ideally, $[L^2]$ should be relatively small, indicating that the optimum falls near the center of the range of percentage concentrations.

With four or more percentages of ingredient, the remaining degrees of freedom test whether the parabola agrees satisfactorily with the observed scores. The total sum of squares for treatments is given by Equation (2), which may be extended if necessary, to include a fifth or sixth concentration. The difference between the total sum of squares for treatments and $[L^2] + [Q^2]$ measures the variation of the mean scores about the curve used in computing the maximum. If its mean square does not exceed the error significantly, the parabola approximates the true relation between x and y. The maximum is then calculated from the full range of observations. If the discrepancy is significant, the concentration for the maximum y may be recomputed from a restricted range of observations, omitting the concentration farthest from the maximum.

Granted the above conditions, Equation (4) can be by-passed and the optimal value computed directly from the sums of the scores as

$$x_{\max} = \bar{x} + x_m \qquad . \ . \ . \ . \ . \ . \ (8)$$

where $\bar{x}$ is the mean or mid-concentration of those used in the calculation and

$$x_m = \frac{I(C-A)}{2(2B-A-C)} \quad \text{for 3 concentrations (9a)}$$

$$x_m = \frac{I(3D+C-B-3A)}{5(B+C-A-D)} \quad \text{for 4 concentrations (9b)}$$

$$x_m = \frac{7\,I(2E+D-B-2A)}{10(B+2C+D-2A-2E)} \quad \text{for 5 concentrations (9c)}$$

The standard error for x_m applies equally to x_{max}. The numerator and denominator in Equations (5) and (9) are uncorrelated, and both are subject to error. The variance of a ratio may be written for Equation (5) as

$$V\left(\frac{b_1}{2b_2}\right) = \frac{b_1^2}{4b_2^2}\left\{\frac{V(b_1)}{b_1^2} + \frac{V(2b_2)}{4b_2^2}\right\}$$

Substituting terms corresponding to those in Equation (9), the standard error of x_m (the square root of its variance) may be reduced algebraically to the form

$$s_{x_m} = |\, x_m \,|\, s^2 \sqrt{\frac{1}{[L^2]} + \frac{1}{[Q^2]}} \;.\;.\;.\;.\;. \quad (10)$$

where the vertical lines enclosing x_m indicate that the sign of x_m is here always positive. This is applicable to tests with three or more concentrations. Another derivation giving exactly the same result for the case of three concentrations only may be written in terms of the original totals as

$$s_{x_m} = \frac{I\sqrt{[(C-A)^2+(B-A)^2+(C-B)^2]Ns^2}}{(2B-A-C)^2} \;.\;.\;.\;. \quad (10a)$$

and used as a check on the arithmetic.

The original data on serum solids are summarized in Table 5 and analyzed in Table 6. The preferred concentration is evidently near 10 percent. The linear and quadratic terms computed from Equations (6) and (7) for four concentrations are $[L^2] = [-3(13.60) + 0.56 - 20.38 + 3(7.34)]^2/(20\times 29) = (-38.60)^2/580 = 2.5689$ and $[Q^2] = [20.38 + 0.56 + 7.34 + 13.60]^2/(4\times 29) = (41.88)^2/116 = 15.1201$, which have been entered in the first two lines of Table 6. These account for two of the three degrees of freedom between the four treatment totals. The sum of squares for all three degrees of freedom computed by Equation (2) is 22.5687, giving by difference 4.8797 with one

TABLE 5.

Frequency distribution of scores in tests on preferred content of serum solids

Serum solids percent	Boys					Girls					Total
	1.03	30	–.30	–1.03	*S(fy)*	1.03	.30	–.30	–1.03	*S(fy)*	*S(fy)*
8	4	4	1	10	–5.28	3	1	1	5	–2.06	–7.34 = A
10	13	6	0	0	15.19	3	7	0	0	5.19	20.38 = B
12	2	6	9	2	–.90	4	1	3	2	1.46	0.56 = C
14	0	3	9	7	–9.01	0	1	6	3	–4.59	–13.60 = D

TABLE 6.

Analysis of variance for data in Table 5

	Degrees of freedom	Sums of squares	Mean square	*F*
Parabola based {linear term, [L^2]	1	2.5689	2.5689	4.80
on 4 concs. }quadratic ", [Q^2]	1	15.1201	15.1201	28.23
Discrepancy from parabola	1	4.8797	4.8797	9.11
Boys vs. girls	3	.7958	.2653	.50
Error	81	43.3877	.5356	
Total	87	66.7522	.7673	
Parabola based {linear term, [L^2]	1	1.0760	1.0760	2.01
on 3 concs. }quadratic ", [Q^2]	1	12.9888	12.9888	24.25

degree of freedom for testing agreement with the parabola. The steps for completing the analysis of variance in Table 6 parallel those detailed in the preceding section.

It is evident from the analysis of variance that boys did not differ from girls in the preferred concentration of serum solids. However, the parabola differed significantly from the total scores at the four concentrations. Because of this discrepancy, it is desirable to narrow the range of concentrations by omitting that furthest from the one receiving the highest score. The linear and quadratic variances for the parabola computed from the totals A, B and C by Equations (6) and (7) have been entered at the foot of Table 6. The non-significant linear term and the highly significant quadratic term indicate a satisfactory basis for calculating the preferred concentration of serum solids from A, B and C. Substituting the numerical values in Equations (9a) and (8), we have

$$x_m = \frac{2(0.56+7.34)}{2(40.76+7.34-0.56)} = .1662 \quad \text{and}$$

$$x_{max} = 10 + .166 = 10.166\%.$$

The error may be estimated from Equation (10) as

$$s_{x_m} = .1662\sqrt{.5356 \quad \left\{\frac{1}{1.0760} + \frac{1}{12.9888}\right\}} = .122$$

However, our confidence in the estimate of x_{max} is lessened by the need of omitting one percentage due to the failure of the parabola to fit all four concentrations within the sampling error. Equation (10) makes no allowance for a discrepancy of this sort. A simple adjustment, which probably over-corrects the error but avoids placing undue confidence in the precision of x_{max} is to multiply s_{x_m} by $\sqrt{F}$ for the discrepancy from the parabola. In the present case $\sqrt{9.11} = 3.02$, giving an adjusted error of 3.02 x 0.122 = 0.37 percent. Hence the preferred concentration of serum solids in ice cream containing 14 percent fat has been determined as 10.17 ± 0.12 percent although the standard error of this estimate may be as large as 0.37 percent.

The scores obtained in the choice tests on the fat content of ice cream are given in Table 7 and analyzed in Table 8.

TABLE 7.

Frequency distribution of scores in tests on preferred fat content of ice cream

Percent	Boys					Girls					Total	
Fat	1.03	.30	–.30	–1.03	*S(fy)*	1.03	.30	–.30	–1.03	*S(fy)*	*S(fy)*	$[y^2]$
8	0	7	20	20	–24.50	1	2	2	6	–5.15	-29.65 = A	16.2519
10	16	16	9	6	12.40	0	1	7	3	–4.89	7.51 = B	24.0472
14	23	15	5	4	22.57	8	3	0	0	9.14	31.71 = C	20.7686
18	8	9	13	17	–10.47	2	5	2	2	.90	-9.57 = D	30.9701

TABLE 8.

Analysis of variance for data in Table 7

		Degrees of freedom	Sum of squares	Mean square	*F*
Parabola based on 4 concs.	linear term, $[L^2]$	1	3.5090	3.5090	6.41
	quadratic ", $[Q^2]$	1	31.5357	31.5357	57.57
Discrepancy from parabola		1	.0006	.0006	.00
Difference between parabolas fitted separately to boys and girls		2	2.2522	1.1261	2.06
Interaction of discrepancy by sex		1	4.1690	4.1690	7.61
Error		168	92.0378	.5478	
Total		174	133.5044	.7673	

Because of the unequal intervals between successive concentrations, the parabola in Equation (4) has been fitted directly by simultaneous equations. The linear and quadratic variances $[L^2]$ and $[Q^2]$ were separated by means of orthogonal coefficients suitable for this case and were of a magnitude allowing good estimation of x_m. In contrast with the experiment on serum solids, the parabola computed from four concentrations agreed excellently with the total scores. Although the parabolas fitted separately to the data for boys and for girls agreed within the limits of error, the observed scores diverged from these parabolas quite differently. The preferred ice cream with 11 percent of serum solids is estimated to contain 13.49 ± 0.23 percent of fat, boys and girls concurring in this result.

Homogeneity of the variance within samples. In the above experiments the transformation to scores presumably has stabilized the variance in the response to each of the four alternatives in a given series. This assumption underlies the equations for the error of the preferred concentration of a given ingredient. However, it would not be unexpected for subjects to show a greater agreement in their preferences for some alternatives than for others quite apart from the statistical factors leading to the transformation from ranks to scores.

This possibility is illustrated in Table 5, where all subjects selected 10 percent of serum solids for either first or second choice but an 8 percent concentration elicited a wide range of opinion, 15 subjects considering it the poorest of all four samples, 7 finding it the best and the other 7 giving it an intermediate ranking. From the frequency distributions of the scores for each sample it is a simple matter to compute the sum of squares of deviations from the separate mean scores for boys and girls. The results from Table 5 are shown in Table 9.

TABLE 9.

Chi-square test of the homogeneity of the variances within sexes for each concentration of serum solids in Table 5

Serum solids percent	Sum of squares = $[y^2]$			log $[y^2]$
	Boys	Girls	Total	
8	13.8353	8.2428	22.0781	1.3440
10	2.1877	1.1191	3.3068	.5194
12	5.5510	6.5122	12.0632	1.0815
14	4.2337	1.7059	5.9396	.7738
Total			43.3877	3.7187
Mean			10.8469	$n = 81$
Log mean			1.0353	$k = 4$

Since the sex of the subject did not modify the relative order either of the mean response or of the sum of squared deviations, the sums of squares for the two sexes may be added for each concentration. The total of the resulting four sums of squares is then equal to 43.3877, the error in the analysis of variance in Table 6. Similarly in Tables 1, 3 and 7, the sum of squares has been computed for each sample and checked against the corresponding terms in the analysis of variance.

The "sums of squares" or $[y^2]$'s measuring the variability among the scores of each sample are then tested for homogeneity. Do they differ from one another more than could be expected by chance? The homogeneity of a set of independent $[y^2]$'s may be determined by computing

$$\chi^2 = \frac{6.9078\, n^2}{3n+k+1} \left\{ \log\overline{[y^2]} - \frac{S(\log [y^2])}{k} \right\} \quad . \; . \; . \; . \; . \quad (12)$$

where $\overline{[y^2]}$ is the arithmetic mean of the k individual $[y^2]$'s representing a total of n degrees of freedom. This modified form of Bartlett's (1) equation is suitable for series where the degrees of freedom in each component $[y^2]$ are equal. If they are also independent, χ^2 computed by Equation (12) is referred to a table of χ^2 (5, 9) with k - 1 degrees of freedom to test whether the variance differs significantly between samples.

In the case of ranked data, however, the $[y^2]$'s for the different preparations in a series are not independent of each other. This is evident from the degrees of freedom (n) in the pooled error. In arranging four objects in order, the position of the fourth object is fixed as soon as three of them have been classified. Hence we have $4 - 1 = 3$ degrees of freedom between the scores of each subject, and with 19 male and 10 female subjects in the experiment of Table 5, there are $(18 \times 3) + (9 \times 3) = 81$ degrees of freedom in the sum of squares for the pooled error. If the variance were required for each type of ice cream, $81/4 = 20.25$ degrees of freedom would be assigned equally to the $[y^2]$ in Table 9 for each of the four concentrations of serum solids, quite a different number from the customary value of $29 - 2 = 27$ degrees of freedom (correcting separately for the means of both sexes). With fewer degrees of freedom in the $[y^2]$ for each concentration than the algebraic sum of the number of squares from which it is computed, the several $[y^2]$'s of a series are not independent.

In view of the correlation between the sums of squares compared by χ^2, Cochran* suggests as a first approximation that the degrees of freedom for testing the significance of χ^2 should be reduced to $k - 2$. The 81 degrees of freedom in the sum of squares for the pooled error in Table 9 would provide three component $[y^2]$'s, each with 27 degrees of freedom, the number expected for independence from a count of the number of squared scores entering any one $[y^2]$. Hence the four $[y^2]$'s represent not more than three degrees of freedom. When they are compared in turn with their mean, yet another degree of freedom is lost, giving us $4 - 2 = 2$ degrees of freedom for the χ^2 for judging the homogeneity of the variances in any of the experimental series reported here.

The computation has been applied to the sums of squares for the four percentages of serum solids in Table 9, converting each $[y^2]$ to logarithms and substituting in Equation (12) to obtain

$$\chi^2 = \frac{6.9078 \times 81^2}{(3 \times 81) + 4 + 1} \left\{ 1.0353 - \frac{3.7187}{4} \right\} = 182.75 \times .1056 = 19.30$$

with $4 - 2 = 2$ degrees of freedom. Heterogeneity as marked as this would not be expected as often as once in 1000 trials. The homogeneity of the variances has been tested similarly for the data in Tables 1, 3 and 7, to obtain χ^2's of 2.17, 1.69 and 4.51, each with two degrees of freedom. Hence in the remaining tests the data were consistent with the assumption of a stable variance between samples.

The tests with serum solids did not conform to this requirement. To estimate the preferred concentration and its error when the variances are unequal is a more involved problem than that described here and has not been considered. However, the present equations provide a first approximation for heterogeneous data. If some of the data is omitted and the variation differs with the concentration, one may prefer to base his estimate of s^2 entirely upon the variation at the concentrations used in computing the preferred level. For the series on serum solids, the total $[y^2]$'s for 8 to 12 percent inclusive

* Personal communication.

with 60.75 "degrees of freedom," gives $s^2 = .6164$, which does not differ materially from $s^2 = .5356$ for all four concentrations. However, it may be used to adjust $s_{\bar{x}_m}$ computed as 0.122 with Equation (10) by multiplying it by $\sqrt{\frac{.6164}{.5356}} = 1.073$ to obtain $s_{\bar{x}_m} = 0.122 \times 1.073 = 0.131$ for the standard error of the preferred concentration of serum solids.

Comparisons of different series. In considering the factors behind consumer preferences for given concentrations of fat or serum solids, the possibility arose that the "body" of the ice cream may have been a determining factor. In this case a subject preferring a lighter ice cream would be expected to react favorably to samples with the lower concentrations of either fat or serum solids and vice versa. Fifteen boys and nine girls participated in both tests and their scores in the two tests have been correlated in examining this hypothesis. The sums of squares and products were computed for an analysis of covariance, pairing the samples of ice cream in the two series in the order of successively increasing concentrations and segregating the effect of concentration and the sex difference from the residual error. The correlation coefficient of the deviations in scores, computed from the row for error, was less than its sampling error. However, when the concentration receiving the highest score in the test of serum solids (10 percent) was matched with that receiving the highest score for fat content (14 percent) and the calculation restricted to these and the two adjoining concentrations (8 percent serum solids with 10 percent fat and 12 percent serum solids with 18 percent fat), the deviations in the error row were significantly correlated. The correlation coefficient $r = 0.36$ with 49 degrees of freedom. Hence we may assume that the "body" of the ice cream helped determine consumer preferences for fat content and for the concentration of serum solids.

Conclusions. This study of consumer preferences, based upon four tests with experimental ice creams, describes quantitative methods suitable for research upon factors such as the palatability, flavor and body of foodstuffs. It also is suitable where sight is the basis of choice, such as in preferences for style of garments and the color, texture and design of fabrics. It is important to select subjects, preferably at random, from the population for which the results are to apply, and for this purpose the principles of stratified sampling (3) are an important guide. The undergraduates in a state university form a relatively homogeneous group so far as age, regional background and economic status are concerned. How widely the food preferences of such a group will apply is problematic.

The "ranking" system of identifying relative quality has the distinct advantage of simplicity over the more familiar method of affixing a "grade." The "grade" fixes the item with respect to an established standard. Considerable training is necessary before an individual can obtain consistent results in grading a series of items. Inexperienced and unbiased consumers, on the other hand, can rank a series of samples from best to poorest on the basis of their own standards. These ratings are often of importance to the manufacturer. By means of the present method a series of items ranked by many people can be evaluated statistically to determine consumer and local preferences.

Each of the present experiments was restricted to four alternatives. There is no statistical restriction, however, to the number which can be handled. The limitation to four choices in each set would rule out an important modification in experimental design, the complex or factorial experiment. In the two series testing the content of fat and of serum solids, for example, we do not know how much the preferred concentration of one ingredient depends upon the level of the other. By increasing the number of samples in a given series to nine, it would be possible, for example, to test the relative preferences for all possible combinations of 8, 10 and 12 percent of serum solids with 10, 14 and 18 percent of fat to determine the extent to which the concentration of one ingredient determines the preferred concentration of the other. For determining the most palatable vanilla flavoring, the importance of testing several concentrations of the natural and of the artificial product in various combinations has already been mentioned.

The efficiency of the technique may be increased by testing each prospective subject on two separate occasions with the same series of samples numbered differently and at random. Suitable subjects are those whose scores for the same samples are positively and significantly correlated. Their agreement or disagreement with others in the test, however, should be disregarded to avoid biassing the results.

Summary. Consumer preferences for two qualitative factors in ice cream, vanilla and chocolate flavoring, and for two quantitative factors, percentage concentration of serum solids and of fat, have been tested on college students. The subjects ranked the four alternative ice creams of each series in order of choice. These ranks were then converted to scores suitable for use in the analysis of variance, with which the significant consumer preferences were identified. American process chocolate was preferred to three Dutch types, which were not scored equally. Natural vanilla flavoring rated higher than the artificial product but a 50-50 mixture of the two scored highest of all. In ice cream containing 14 percent fat, the preferred concentration of serum solids was 10.17 ± 0.37 percent; in that containing 11 percent of serum solids, the preferred concentration of fat was 13.49 ± 0.23 percent. Both sexes concurred in these results.

The design and statistical analysis of choice tests are given in detail. Simple methods are described for computing the concentration giving the maximum score and its error and for testing the homogeneity of the response to different items in a series. Possible applications of the technique are indicated.

ACKNOWLEDGMENTS

The authors are indebted to Professors R. A. Fisher, W. G. Cochran and B. H. Camp for advice concerning the error of the preferred concentration.

REFERENCES

(1) Bartlett, M. S. Some examples of statistical methods of research in agriculture and applied biology. Suppl. Jour. Roy. Stat. Society. *4*: 137-170. 1937.

(2) Cobb, J. S. A study of culinary quality in white potatoes. Potato Jour. *12*: 335-346. 1935.

(3) Cochran, W. G. The use of the analysis of variance in enumeration by sampling. Jour. Am. Stat. Assoc. *34*: 492-510. 1939.

(4) Dove, W. Franklin. The relative nature of human preference: with an example in the palatability of different varieties of sweet corn. Jour. Compar. Psych. *35*: 219-226. 1943.

(5) Fisher, R. A., and Yates, F. Statistical Tables for Biological, Agricultural and Medical Research. Oliver and Boyd, London. 1938.

(6) Friedman, Milton. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Jour. Am. Stat. Assoc. *32*: 675-701. 1937.

(7) Hotelling, Harold. Experimental determination of the maximum of a function. Ann. Math. Stat. *12*: 20-45. 1941.

(8) Kendall, M. G., and Smith, B. Babington. The problem of *m* rankings. Ann. Math. Stat. *10*: 275-287. 1939.

(9) Snedecor, G. W. Statistical Methods. Iowa State College Press. 1940.

(10) Stevenson, F. J., and Whitman, E. F. Cooking quality of certain potato varieties as influenced by environment. Am. Potato Jour. *12*: 41-47. 1935.

(11) Trout, G. M., Downs, P. A., Mack, M. J., Fouts, E. L. and Babcock, C. J. The evaluation of flavor defects of butter, cheese, milk and ice cream as designated by dairy products judges. Jour. Dairy Sci. *25*: 557-569. 1942.

(12) Trout, G. M., Downs, P. A., Mack, M. J., Fouts, E. L., and Babcock, C. J. Comparative standardization of butter, cheese, milk and ice cream flavor scoring. Jour. Dairy Sci. *26*: 63-68. 1943.

(13) White, William, Downs, P. A., Mack, M. J., Fouts, E. L. and Trout, G. M. Correlation between grades on scores and grades on criticisms in the judging of dairy products. Jour. Dairy Sci. *23*: 1-12. 1940.