

Measurement of Sensory Differences^a

DAVID R. PERYAM AND VENONA W. SWARTZ
Quartermaster Food and Container Institute for the Armed Forces, Chicago, Illinois

(Received for publication, January 2, 1950)

There is great need for objectivity in the field of flavor measurement. Relatively greater objectivity is possible by use of tests which depend upon discrimination rather than upon judgment. Three such tests designed for the measurement of sensory differences are described, and a method for statistical analysis of the results is suggested.

Quality, as determined by the ultimate crucial test of consumer acceptance, is paramount to every food processor, and, concomitantly, to all those who are engaged in food research and development. Consumer acceptance of a food product is conditioned largely upon its flavor quality, i.e., all those properties affecting human senses of taste and odor which determine its pleasantness or unpleasantness. Other aspects of a food are important in determining its total worth, such as nutritional adequacy, microbiological purity, and chemical stability; but without satisfactory flavor quality, it may not matter that a food is otherwise good, for a food product which is adequate in every other way may be rejected by individuals and by whole sections of the population simply because "it doesn't taste good."

These secondary aspects, if we may call them that, are relatively easy to deal with. Ordinarily prediction may be made and control effected in these areas by standard scientific techniques if proper equipment and trained personnel are available. But in the crucial aspect of flavor quality, the *sine qua non* of consumer acceptance, we encounter a more difficult situation. Food technologists today are becoming aware of the problem but few face it in a realistic manner. Attempts to deal directly with flavor problems have generally been variable and not characterized by precision. This results from the fact that flavor problems as a whole belong to a field which is diffuse and many-sided, while individual problems may appear deceptively easy and hence become dangerously subject to oversimplification. Add to this the fact that there exists no standardized methodology in this field and it is understandable why many food processors choose to trust solution of their flavor problems to luck, aided perhaps by careful attention to process variables, rather than to subject them to direct attack.

Why is there this tendency toward oversimplification in a complex field, this lack of appreciation of the true difficulties involved? While everyone is aware that a vitamin assay, to take an example, requires both special equipment and a special knowledge on the part of the

^a This paper reports research undertaken by the Quartermaster Food and Container Institute for the Armed Forces, and has been assigned Number 289 in the series of papers approved for publication. The views or conclusions contained in this report are those of the authors. They are not to be construed as necessarily reflecting the views or indorsement of the Department of the Army.

analyst, when a problem of flavor evaluation arises the researcher may reason like this, "What equipment is required? Why, simply normal human senses such as are available to me and to everyone else. What know-how is required? All one has to know is what he has tasted and how to describe it. Anyone can do that if he only stops to think. So the boys and I will just sit down and taste it ourselves." Such a simple, naive philosophy is not far removed from the flavor control method that is in most common use, i.e., the small panel technique using a score card with an arbitrary scale. Even though it is basically naive, the panel technique can be worthwhile if used properly and in some situations it is the only approach possible. But it is used indiscriminately and in many cases it results in great sacrifice of precision or even in utter failure to obtain useful information. Quality measurement has a world of possibilities not revealed in this simple philosophy.

When predictions fail in the realm of flavor quality there is a tendency to say: "People differ. Their sensory capacities differ. Also, there is no accounting for tastes. You can't expect precision when you are dealing in human behavior." This is but apology. People are also much the same, their sensory capacities are a biological fact, and even their preferences and their judgments are subject to quantitative description. We must recognize stability as well as variability. Science measures in such a way that the measurements can be used for prediction and control. Human behavior can be dealt with scientifically, just as can the subject matter of physiology, physics, or chemistry. However, knowledge of the techniques required is not wide-spread. This is particularly unfortunate for food technology because of the vital problem of flavor which must, by its very nature, be dealt with in terms of human behavior.

The great need in the field of flavor measurement is objectivity,—test methods which will permit the interpretation of results at a level higher than that of pure speculation. Too often many variables are allowed to operate at once, as for example, in the typical score card system where a number of factors, each measured subjectively by judges who we only hope are all doing the same thing, are converged into a single scale value. The result is likely to be a meaningless artifact, accepted in resignation because there is no better. The constant aim should be to reduce as far as possible the number of variables simultaneously operating, in a manner analogous to the way control is effected in the physical sciences. This not only avoids equivocation in interpretation but has the corollary effect of making the problem for the individual observer less complex.

A test should set a critical situation in which units of human response can be counted and handled statistically. We are on firmer ground when we deal thus with discrimination rather than with judgments. Ad-

R50-8
TECHNICAL SERVICES
NATIONAL BUREAU OF STANDARDS
NATIONAL LABORATORIES
NATIONAL BUREAU OF STANDARDS
NATIONAL BUREAU OF STANDARDS
NATIONAL BUREAU OF STANDARDS

mittedly it is not always possible to avoid the psychologically complex activity implied by the term "judgment," nor would we want to do so, for that activity has many proper uses, such as the rapid estimation of pleasantness value of foods or the measurement of purely subjective factors. But there arise many problems which may be attacked directly by what we may call discrimination tests. One such problem is the determination of the existence of a difference between foods which is capable of detection by sensory means. This problem is particularly easy to solve by objective testing.

This article presents three tests designed for difference testing, all of which fulfill the criteria of objectivity and ease of interpretation. They are variants of one basic type, but each of them has certain advantages. Each sets a definite discrimination problem in such a way that the observer can give a positive straightforward answer, and results in each case can be interpreted statistically.

The general concept of difference testing based upon discrimination rather than judgment is not a new one. Methods suggestive of that principle have been described in the food research literature from time to time but the principle has never been carefully formulated. The methods herein described were worked out in the Quality Research Laboratory of Joseph E. Seagram and Sons in Louisville, Kentucky, in 1941 and 1942, and since then that company has been using them with marked success in both quality research and production control. The use of these methods in certain problems of product analysis has been described (5) by the founder of the Seagram Laboratory, Dr. E. H. Scofield, who was the major contributor to their development. One of the three, the triangular test, was developed independently of Seagram's work, in the research laboratory of the Carlsberg Breweries, Copenhagen, Denmark, where it has been used for control work and for selection of taste panels. The form of the test used there is the same but the test conditions do not appear to be as well controlled. The work was published in the United States in 1946 (3) and the method is sometimes referred to as the "Helm technique." In 1936 (1) and again in 1940 (2) Cover published articles describing a test procedure which she named the "paired eating method" and which she employed to measure differences in regard to tenderness and other qualities in meats. It is an objective method which requires meticulous control of most test variables and which produces data capable of statistical verification. It is quite similar to the duo-trio test of the present paper except that it does not use a reference standard.

The Food Acceptance Laboratory of the Quartermaster Food and Container Institute has recently adapted these efficient sensory research tools, in the forms herein described, to a wide variety of food research problems where their utility has been further demonstrated.

The Duo-Trio Test

Of two products which are to be tested for difference one is selected as the "control." Logically, either of the two may be used, but if either is more familiar to the

subjects or if either is more likely to be considered as normal, that one should be selected. For example, when checking a stored product against the fresh product for possible deterioration, always use the latter as control. Uniform samples of the control and of the unknown being tested are prepared and are presented to the observer in succession. The observers should be thoroughly familiar with the test and should know the entire sequence of samples, except, of course, the order of the critical pair. These are the basic instructions for the test:

"First you will be given a warm-up sample. This will be the same as the control sample which follows but be prepared to disregard it, for it is apt to taste different merely because it is first. It is given to you only to get the flavor in your mouth. Rinse your mouth with the water which is provided after this sample and also after each of the samples which follow. Next you will get a control, and following that, two samples in unknown order. One of them will be the same as the control but the other will be different. It is your problem to say which one of the two, whether the first or the second, is the different one."

The observer is required to give a decision, being instructed to make a "best guess" if he is not sure. The reason for forcing a judgment will be apparent when we consider the method of summarizing results. We have found that with most food substances tested two of these basic test units can be presented at a single sitting without apparent effect on the precision of the results. The rinse water should be of the most neutral character obtainable. The distilled water which is available in some laboratories may not be as good for this purpose as tap water, but, in any event, the emphasis should be on low intensity of flavor. The rinse water should be the same temperature as the test samples when tests are run at room temperature. When testing heated samples, the rinse water should be at, or slightly above, normal body temperature. The routine of a single sitting for a subject is diagrammed in Figure 1.

It is essential that certain variables be rigidly controlled in order to assure that accidental differences do not affect the results and that maximum precision is attained. The samples as presented to the observer must be exactly alike in respect to all factors under ex-

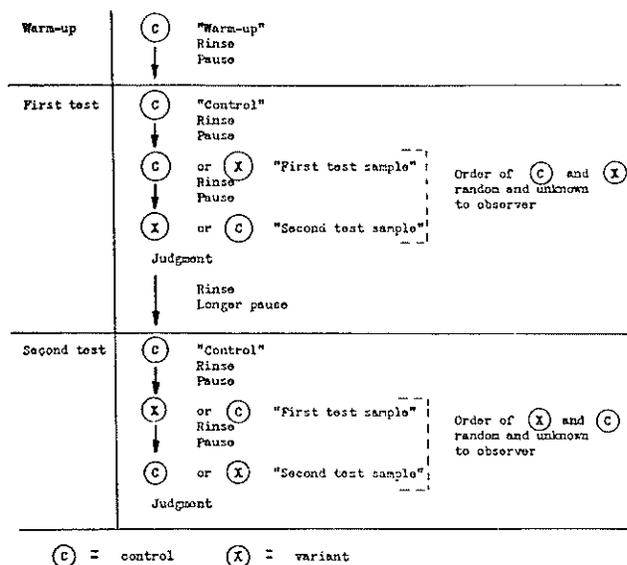


FIG. 1. Diagram of procedure of duo-trio taste test.

perimental control. They must all be at the same temperature, must be presented in identical containers, and must be of the same size. To insure effectiveness of quantity control the amount must be no larger than can comfortably be taken into the mouth at one time and observers should be instructed to handle them in this manner. The products must present the same visual appearance. Although differences in appearance can also be tested by this method, it must be done in a separate test. Determination of the order of the samples within the unknown pairs should be made by some method which insures randomness. If this is not done, the observers may try to "outguess" the experimenter, either consciously or unconsciously, and when this happens, the results bear little relation to the object of the test.

To insure efficiency on the part of the individual observer, the factors of adaptation and mutual interference between samples must be controlled. The "warm-up" and the rinse between samples are both for this purpose. Also, a standard framework of test procedure must be maintained. The time interval between samples is a critical point. Two factors, adaptation and forgetting, work in opposition. On the one hand the time interval should be extended as far as possible to permit sensory recovery, but, on the other hand, it must not be too long because forgetting becomes increasingly important as the interval is extended. The best evidence indicates that the interval should not be less than 10 seconds nor greater than 30 seconds. It may be varied according to the degree of adaptation induced by the substances tested and according to the skill of the observers.

The practice of having each observer make two judgments is a matter of convenience only. For example, the test would be meaningful whether 20 observers gave one response each or one observer repeated 20 times. However, it should be kept in mind that the method measures differences only as perceived and responded to by the observers actually participating. Both skill in the testing situation and sensitivity will vary a great deal, not only from one person to another, but to a lesser extent from time to time with a single person. It is less significant to establish that one or two exceptional individuals can reliably distinguish between two products than it is to show that a larger group will detect the difference. In using a limited number of observers oversensitivity is seldom a problem. The great danger is that the observers may be operating below the optimum and thus fail to detect a substantial difference. The best solution is to have a group selected on the basis of their sensitivity and their skill on this specific procedure and to use a large enough number on any one test to level out chance fluctuations in individual performance. Then one can be reasonably certain that difference ratings are not only comparable but also meaningful.

Tabulation of the test data will give a certain number of correct identifications which can be stated as a percentage of the total judgments. From the test situation itself, which requires guessing in case of doubt, it can readily be seen that when there is no difference between samples chance alone should result in about 50 percent correct identifications. (Here "correct" means merely that the observer chooses the test sample rather than the control.) It can be shown theoretically, and verified experimentally, that the testing of identical samples will result most often in 50 percent "correct" responses,

a situation analogous to getting 10 heads and 10 tails in 20 tosses of a coin, and in higher or lower percentages with continually decreasing frequencies. The probability of obtaining any given percentage correct in this situation can be determined by any one of several methods. When a difference exists between the products tested, however, some of the observers will detect it and their responses will no longer be guesses but will tend to be right more often than chance alone would allow. As the degree of difference becomes larger, more observers will detect it or will detect it more often and the percentage of correct judgments will increase correspondingly. To determine the significance of any given result the problem becomes that of determining how often one would expect to obtain that percentage correct had the samples been identical. The smaller this expectancy, the higher the level of confidence one can have in the result, i.e., the greater the certainty that a true difference exists between the products. One fairly simple method of calculating this expectancy is presented here.

A statistic known as the Critical Ratio is derived. This is the ratio of the difference between the two percentages (the percent correct actually obtained and the theoretical 50 percent "pure-chance" result) to the Standard Error of the latter percentage. The applicable formulae are:

$$C.R. = \frac{P_{obs} - p}{\sigma_p} \quad (1)$$

$$\sigma_p = \sqrt{\frac{p(1-p)}{N}} \quad (2)$$

where C.R. = Critical Ratio; p_{obs} = observed percentage (or proportion) correct; p = percentage (or proportion) correct expected by chance; σ_p = standard error of p ; and N = total judgments.

By solving equation (2) for $p = 0.50$ and substituting in equation (1), the latter can be reduced to simple form, as follows:

$$C.R. = \frac{P_{obs} - 0.50}{\sqrt{\frac{0.50 \times 0.50}{N}}} = \frac{(P_{obs} - 0.50) \sqrt{N}}{0.50} \quad (3)$$

Figure 2 shows a typical data sheet with the Critical Ratio calculated using formula (3). The significance of the result is determined by using the tables of areas under the normal curve with the C.R. figure as a positive sigma distance. The tables will give the proportion of the area lying below this point and by subtracting this from 0.50 one gets the proportion lying beyond. This is also the expectancy, in terms of chances in 100, that identical samples would have given the same result; hence, the smaller this figure becomes, the greater is the confidence one can have that a true difference is represented. This is the concept usually referred to as "confidence level" or "level of significance." The data of Figure 2 give a confidence level just slightly

Date: December 22, 1949

DUO-TRIO DIFFERENCE TEST

Sample No. 1 ----- (Control)
 Sample No. 2 ----- (Variant)

Observer	Order of Unknowns		Judgment		X = Correct judgment O = Incorrect judgment
	1st	2nd	1st	2nd	
DG	1-2	1-2	X	X	Total judgments = 20 Number correct = 15 Percent correct = 75% $C.R. = \frac{.75 - .50}{.50} \sqrt{\frac{.20}{.50}} = 2.25$.50 Significant at 1.3% level of confidence
CB	1-2	2-1	X	O	
SB	2-1	1-2	X	X	
DP	2-1	2-1	O	O	
CH	2-1	1-2	X	X	
MS	1-2	2-1	O	X	
ML	1-2	1-2	X	X	
RS	2-1	2-1	X	O	
AE	1-2	2-1	X	X	
ET	2-1	1-2	X	X	

FIG. 2. Typical data sheet for the duo-trio test, showing statistical summary.

poorer than 1 percent, which is considered very significant.

There are several methods available for analysis of these data. Helm and Trolle (3) used the chi square method which, with 20 or more judgments, will give results identical with those obtained by the Critical Ratio method described here. Also, exact probabilities can be calculated by the binomial expansion theorem. This latter method involves considerable labor and is probably not worth while when dealing with larger numbers of judgments since approximations obtained by the other methods are very close when N is 20 or more. However as N is reduced the error arising in both the chi square and the Critical Ratio methods becomes progressively larger and neither should be used when N is less than 16. Below this point the exact probabilities should be calculated by the binomial theorem.

The Triangular Test

This test presents to the observer a problem which, psychologically, is somewhat more complex than that of the duo-trio test, and in which controlled conditions are not as easily maintained; however, it has been found to give greater precision in many instances. It need not be considered as merely a short-cut, although it does have the added advantage of being easier to conduct and requiring less of the experimenter's time. Also, when a readily detectable difference exists, fewer judgments are required to attain statistically significant results.

A control is used, although it need not be so designated, and the same recommendation that the more familiar of the two samples be used as control applies here also. The critical samples are a trio of unknowns, two controls and a variant, presented simultaneously. The instruction to the observer when this method is used in a taste test is as follows:

"You will be given a warm-up, after which you will rinse. Then you will be given three samples at the same time. Two of these are identical and one is different. You are to pick out the different one. You should proceed slowly, take as nearly as possible the same amount of sample in each taste, and rinse and pause after each taste, in order to avoid interference between samples."

If the substance under test is not so intense in flavor that recovery from adaptation will be slow, a second "triangle" may be given after a relatively short rest, so that we can diagram the procedure of one sitting as shown in Figure 3.

Here control of amount tasted, and control of adaptation by use of proper time intervals and rinses, must

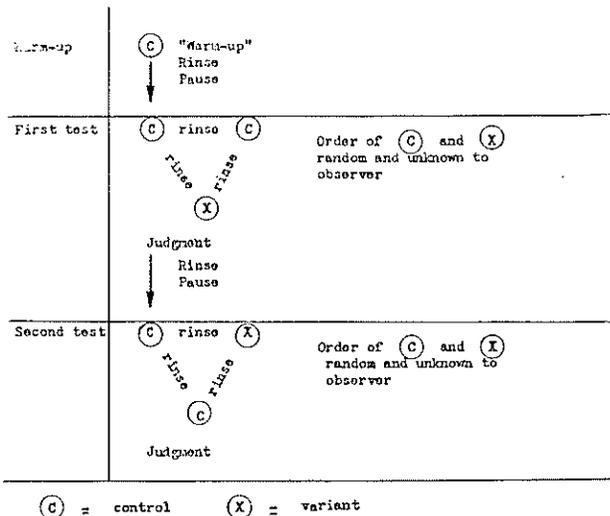


FIG. 3. Diagram of procedure in the triangular taste test.

be the responsibility of the individual observer. For this reason a more thorough training is required if the test is to have maximum precision. The total quantity of sample provided should be such that the observer can check back two or three times if he wishes.

Statistically the data are analyzed just as with the duo-trio test, the only change being that here the expected percentage of correct judgments, if all samples were identical, would be 33-1/3 percent since the observer has only one chance in three of guessing the correct sample. Solving equation (2) and equation (1), as before, using $p = 0.33$:

$$C.R. = \frac{p_{obs} - 0.33}{\sqrt{\frac{0.67 \times 0.33}{N}}} = \frac{(p_{obs} - 0.33) \sqrt{N}}{0.47} \quad (4)$$

For example, if 12 correct judgments are obtained in a total of 16 responses, we have:

$$C.R. = \frac{(0.75 - 0.33) \times \sqrt{16}}{0.47} = 3.6$$

A slightly different method of analysis of triangular test data, but one which yields identical probability values, is described by Roessler, Warren, and Guyman (4).

The Dual-Standard Test

This test is substantially the duo-trio test adapted for use with odor samples. Taking advantage of the fact that recovery from odor stimulation is much more rapid than is recovery from taste stimulation, the observer is provided with a pair of standards, the control and the variant, rather than just the control presented singly, for study prior to presentation of the unknown pair. This gives him an opportunity to develop a more definite criterion of difference, and precision is bettered as a consequence. This basic instruction is given to the subject:

"Here are two odor samples. Note that one is marked "S1" and the other is marked "S2." You are to study these. Note any differences, smelling back and forth until you believe you can tell them apart. Then you will be given this second pair of samples. They are the same as the first but are unidentified.

It will be your problem to decide which one of them is like S1 and which is like S2. Do not hurry. Smell the samples of the pair alternately and pause four or five seconds between sniffs. It is suggested that you smell each sample no more than three times. You may check back on the standards if you wish."

The procedure is shown in diagram form in Figure 4.

This test should be conducted in a room where the air is free from definite extraneous odors. The samples should be held in closed containers: 200-ml. Erlenmeyer flasks with ground-jointed stoppers are very convenient for this purpose. The flasks should be opened only while actually being smelled. The untrained observer will have a tendency to alternate too rapidly between the samples on this test so the part of the instructions relating to the time interval must be particularly emphasized during training. If strong odors are being tested even more time should be allowed between sniffs. When dealing with relatively weak odors such as are normally encountered with food products, an observer can give even more than two responses at a single sitting if a recovery period of about one minute is given after each judgment.

Analysis of results on this test is done exactly as for the duo-trio test.

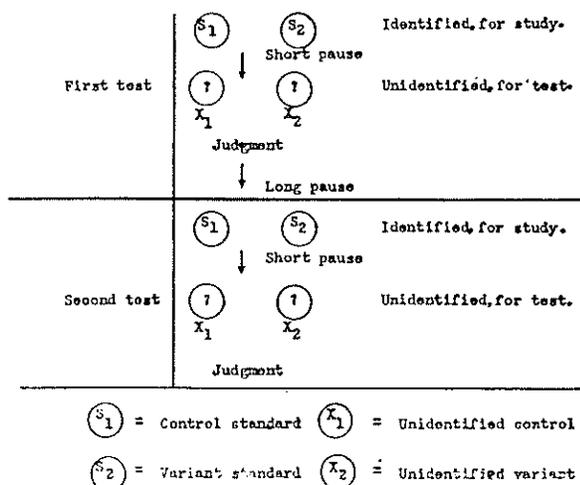


FIG. 4. Diagram of procedure in the dual-standard odor test.

Discussion

Theoretically, any one of these forms could be used for any type of testing: taste, odor, or appearance. Actually, in many instances one form will serve equally as well as another so that the experimenter may choose the one which is most convenient or which can be handled with greatest ease by the observers available. However, there are certain limitations and advantages of each of the three forms which should be kept in mind.

1. *The Duo-Trio Test.* This form has an advantage for taste tests in that it permits more definite control of the observer as, for example, in the matter of time interval and quantity of sample. Its superiority is limited to taste testing since successive presentation of samples is unnecessary and even confusing in odor and appearance tests. It requires careful attention on the part of the laboratory operator. When a difference is fairly evident so that guessing is infrequent, more judgments are required to establish that difference at a given level of confidence than with the triangular test. But with trained operators and a smoothly working laboratory it can give maximum precision.

One example of consistently good results with this method in the Food Acceptance Laboratory has been in the investigation of dried whole milk—tested in its reconstituted form. It has been used first to select a panel of observers of superior sensitivity and then to measure, in terms of this panel's skill, differences caused by deterioration and by processing variables. The flavor of milk seems to be well suited to discrimination in this highly controlled testing situation.

2. *The Triangular Test.* As pointed out before, when this form is used for taste, it is more difficult to maintain experimental control than in the duo-trio, since the responsibility for timing, rinsing, and quantity taken at one time lies with the observer. However, training can overcome this disadvantage and the method in many instances gives more precise results than the duo-trio. Even though the situation, with three unknowns simultaneously under consideration, is psychologically more complex, it definitely helps the subject to remember the samples when he can place them in a visual-spatial frame of reference and also can check back on previous impressions. This factor seems particularly advantageous in testing complex flavors while the more controlled duo-trio is superior when simpler flavors are compared. This form can be used for odor tests, where it is faster than the dual-standard form but it does not give as precise results. It is probably the best of the three for tests involving differences in appearance.

The triangular test has been used in the Food Acceptance Laboratory for a wide variety of difference problems. A good example is that of the evaluation of imitation peppers in terms of how closely they approximate natural peppers in strength and quality. Weak water infusions of both peppers are made up by a standard procedure and are presented in the triangular test to trained observers using the natural spice as the control. Tomato juice seasoned at normal levels is also used as a testing medium. The triangular test has been shown to give better discrimination between peppers than the duo-trio when using either the water infusions or tomato juice.

3. *The Dual-Standard Test.* This is the best form for odor testing because the second standard permits the observer to form a more stable criterion of what to look for in the unknowns. In taste testing, however, presentation of the second standard is a disadvantage since both forgetting and adaptation are enhanced by the lengthened series. An illustration of a problem in which the dual-standard odor test was particularly applicable was the possible effect of various fumigants on stored spices. Standard water infusions were made of the spices that had been in the fumigated area and of controls which were untreated but otherwise matched with the fumigated samples. The control and treated sample were then tested for odor difference by the dual-standard method.

All three forms test only for differences, as such, and are not designed to give information about preference or superiority. This should not be considered a limitation for it arises from the fact that we have purposefully isolated a single factor for experimentation. We can be certain that results are more valid for this very reason.

Once the basic fact of difference has been established, one can test further by other means to determine the corollary effects of that difference.

A word of caution about the results of the statistical analysis is in order. The Critical Ratio figures will look like difference scores. If Product A differs from the control by 2.0 C.R.'s and Product B from the same control by 2.7 C.R.'s, there will be a strong tendency to assert that Product B differs by a greater amount. Actually this may be so, but, statistically, we have no information on the matter; we can only say that we have established the "B"-difference to a higher degree of certainty. The validity of using these C.R. figures as actual difference scores is being tested in the Food Acceptance Research Laboratory at the present time.

LITERATURE CITED

1. COVER, SYLVIA. A new subjective method of testing tenderness in meat. *Food Research* 1, 287 (1936).
2. COVER, SYLVIA. Some modifications of the paired eating method in meat cookery research. *Food Research* 5, 379 (1940).
3. HELM, ERICK, AND TROLLE, BIRGER. Selection of a taste panel. *Wallerstein Lab. Commun.* 9 (28), 181 (1946).
4. ROESSLER, E. B., WARREN, J., AND GUYMAN, J. F. Significance in triangular taste tests. *Food Research* 13, 503 (1948).
5. SCOFIELD, E. H. Some approaches to the measurement of taste and related properties of food and beverages. Conference course lectures on modernizing management methods in the restaurant industry. The University of Chicago and the National Restaurant Association. Pamphlet No. 25 (1948).