AN APPLICATION OF THE CONTROL CHART METHOD TO THE TESTING AND MARKETING OF FOODS*

By Sophie Marcuse

Bureau of Human Nutrition and Home Economics
U. S. Department of Agriculture

The control chart method was developed by Shewhart and others[1,2] for controlling quality in industrial mass production. It is the purpose of this paper to show that this method can also be used to analyze test results in the field of experimental food research.

The ultimate purpose of food research is to improve the eating quality and the marketability of foods. In order to make foods more agreeable to the taste, sound principles must be developed for measuring and interpreting data on their palatability.[3] Although more and more physical and chemical criteria are being devised for testing foods, the human senses of taste and smell remain the most important factors in determining the flavor, texture, aroma, and odor of food products. Testing food products for the subjective reaction to them is known as organoleptic testing.

Two methods commonly used for evaluating a food product organoleptically are: (1) A consumer-preference test is made to find out how well the public likes a given food; (2) A tasting panel, consisting of a limited number of specially trained tasters, is selected to determine how nearly a given sample comes to some pre-established standard.[4] In testing butter, for example, the judges are schooled in scoring according to a previously fixed standard. The present paper discusses the application of control charts to the second method of organoleptic testing, namely, food testing by a panel of tasters.

In such panel testing the usual procedure is to submit control or check samples along with the experimental samples. The fact that the

* A paper presented at the 104th Annual Meeting of the American Statistical Association, Washington, D. C., December 29, 1944.

Acknowledgment is expressed to Dr. W. Edwards Deming for his very kind help in preparing this paper.

[1] W. A. Shewhart, *Economic Control of Quality of Manufactured Product*, D. VanNostrand and Co., 1931; *Statistical Method from the Viewpoint of Quality Control*, ed. by W. E. Deming, The Graduate School, U. S. Department of Agriculture, Washington, D. C., 1939; Leslie E. Simon, *An Engineers' Manual of Statistical Methods*, John Wiley and Sons, 1941.

[2] W. Edwards Deming, "Opportunities in Mathematical Statistics, with Special Reference to Sampling and Quality Control," *Science*, Vol. 97, March 5, 1943.

[3] M. D. Sweetman, "The Scientific Study of the Palatability of Food," *Journal of Home Economics*, Vol. 23, No. 2, February 1931.

[4] Washington Platt, "Some Fundamental Assumptions Pertaining to the Judgment of Food Flavors," *Food Research*, Vol. 2, No. 3, 1937.

check samples are all of the same quality is not known to the tasters. For illustrative purposes, let us assume that a panel consisting of three tasters has scored the palatability of such check samples using a scale of 1 to 10. Let us furthermore assume that the tasting was repeated 4 times a week over a period of 7 weeks. In the table the hypothetical scores of the tasting panel are tabulated. The scoring, as might be expected, shows great variation.

PALATABILITY SCORES OF THREE TASTERS ON CHECK SAMPLES OF SAME QUALITY—HYPOTHETICAL DATA

	Test scores*				Average ($\bar{X}$)	Standard deviation (σ)
	1st	2nd	3rd	4th		
1st week						
Judge A	7.0	7.0	7.0	7.5	7.1	.22
B	6.0	6.0	6.0	6.5	6.1	.22
C	6.5	5.0	4.5	5.0	5.2	.75
2nd week						
Judge A	5.0	6.0	5.5	6.0	5.6	.41
B	5.0	5.0	5.0	5.5	5.1	.22
C	5.5	5.0	4.5	5.5	5.1	.41
3rd week						
Judge A	7.0	6.5	7.5	7.5	7.1	.41
B	5.0	6.0	4.5	6.0	5.4	.65
C	7.0	7.5	6.5	7.5	7.1	.41
4th week						
Judge A	7.0	6.5	6.5	6.5	6.6	.22
B	5.5	6.5	5.5	6.5	6.0	.50
C	6.0	7.0	5.5	5.0	5.9	.74
5th week						
Judge A	5.5	6.5	6.5	6.5	6.2	.43
B	5.5	5.5	5.0	5.5	5.4	.22
C	4.0	5.0	4.5	4.0	4.4	.41
6th week						
Judge A	5.5	6.0	6.0	6.5	6.0	.35
B	5.0	5.0	5.5	5.5	5.2	.25
C	5.0	6.0	6.5	4.0	5.4	.96
7th week						
Judge A	7.5	6.5	7.0	7.0	7.0	.35
B	5.5	6.5	6.5	5.5	6.0	.50
C	7.0	6.5	7.0	6.5	6.8	.25

* Maximum score 10.0; Minimum score 1.0.

To analyze this variability in scoring, tests of significance are usually applied for determining whether the disagreement among the tasters is statistically significant. However, in taste scoring according to a pre-established standard, an analysis of the validity and reproduci-

bility of each taster's scoring should precede the pooling of the scores. Even if the test of significance indicates that the differences between the tasters' scores are statistically insignificant, this would not imply that either the individual tasters' scores or their pooled averages yield a valid judgment. For example, all the tasters may have consistently scored too high, in relation to the pre-established standard.

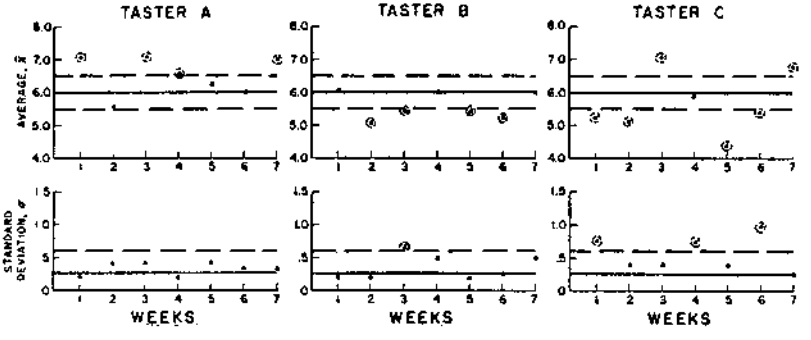
The control chart method can aid in the selection of individual tasters who make the most valid and stable judgments. At the same time this method enables the experimenter to evaluate the tasting results continuously. It therefore can aid producers of food by minimizing the losses arising from two kinds of errors: (a) Failing to recognize a good judge, and thus losing the benefit of valid scores; and (b) failing to recognize and exclude poor judges, thus being guided by invalid scores.[5]

USE OF THE CONTROL CHART METHOD IN TASTE PANEL SCORING

Description of control charts. Charts I and II show a series of graphs resulting from the application of the control chart method to the hypothetical data on organoleptic testing given in the table. Control

CHART I

ANALYSIS OF PALATABILITY SCORES* BY CONTROL CHARTS FOR $\bar{X}$ AND σ, USING SPECIFIED STANDARD VALUES



* Data are given in the table.
Maximum score 10.0; minimum score 1.0.

charts for averages and standard deviations were constructed using the notation of the American Standards Association.[6,7] Each individual taster's scores were analyzed separately. Since a weekly basis was

[5] Deming, *op. cit.*, p. 211.

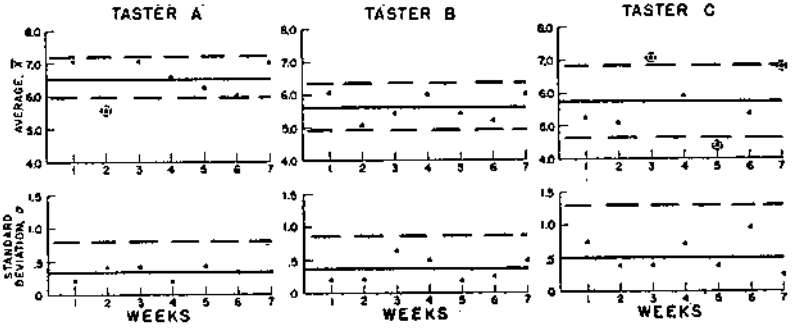
[6] American Standards Association, *Control Chart Method of Analyzing Data*, Z1.2—1941 (American Standards Association, 29 West 39 Street, New York).

[7] American Standards Association, *Control Chart Method of Controlling Quality During Production*, Z1.3—1942 (American Standards Association, 29 West 39 Street, New York).

assumed to be appropriate for this experiment, the data were grouped into weekly rational subgroups combining 4 tastings. The ordinates are the same in Charts I and II; they correspond to the weekly periods. The points on these ordinates show the average score of the week ($\overline{X}$), and the standard deviation of the scores within a week (σ).

CHART II

ANALYSIS OF PALATABILITY SCORES* BY CONTROL CHARTS FOR $\bar{X}$ AND σ, USING STANDARD VALUES DERIVED FROM DATA



* Data are given in the table.
Maximum score 10.0; minimum score 1.0.

In order to set up control charts for $\overline{X}$ and σ, a selection of the standard values $\overline{X}'$ and σ' has to be made first. They serve for computing the central lines and control limits as follows:

Central lines	Control limits
For $\overline{X}$: $\overline{X}'$	For $\overline{X}$: $\overline{X}' \pm A\sigma'$
For σ: $c_2\sigma'$	For σ: $B_1\sigma'$ and $B_2\sigma'$

Since $n=4$, i.e., number of tests per week (or subgroup), the following factors apply:[8]

$$A = 1.500$$
$$B_1 = 0$$
$$B_2 = 1.859$$
$$c_2 = .7979$$

In the case of the hypothetical scores given in the table, it was assumed that no preliminary data were at hand. Two approaches were used in setting up standard values. In Chart I, specified standard values

[8] Computations are based on 3-sigma limits tabulated in appendix I, p. 39, of the American Standards Association, Z1.3—1942 (footnote 7 *supra*).

$\overline{X}'$ and σ' were selected on the basis of engineering judgment; in Chart II, the data themselves were used as a basis for deriving the standard values.

For Chart I, the standard value of the average $\overline{X}$ is specified as 6.0, which is assumed to be the pre-established standard according to which the tasters grade the check samples. This standard value appears as the central line of the three control charts for $\overline{X}$ in Chart I. If the standard value for the standard deviation is specified as $\sigma' = .33$, then the control limits for $\overline{X}$ are:

$$\overline{X}' \pm A\sigma' = 6.0 \pm 1.500 \text{ times } .33 = 6.5 \text{ and } 5.5.$$

Thus the distance between the lower and upper control limits of the weekly averages is one unit in the scoring scale. Observations of previous judgings in related problems indicate this to be a reasonable limit. On the basis of the standard values just adopted, the central line for σ is $c_2\sigma' = .27$, and the control limits for σ are $B_1\sigma'$ and $B_2\sigma' = 0$ and .62.

For Chart II, the standard values $\overline{X}'$ and σ' were adopted as $\overline{X}' = \overline{\overline{X}}$ where $\overline{\overline{X}}$ is the average of all the subgroup values of $\overline{X}$, and $\sigma' = \bar{\sigma}/c_2$ where $\bar{\sigma}$ is the average of all the subgroup values of σ. The central lines and control limits for the three tasters are, therefore, as follows:

	Central lines	Control limits
Taster A	For $\overline{X}$: 6.5	For $\overline{X}$: 7.2 and 5.9
	For σ: .34	For σ: 0 and .80
Taster B	For $\overline{X}$: 5.6	For $\overline{X}$: 6.3 and 4.9
	For σ: .36	For σ: 0 and .85
Taster C	For $\overline{X}$: 5.7	For $\overline{X}$: 6.8 and 4.6
	For σ: .56	For σ: 0 and 1.31

Interpretation of results. In interpreting the tasting scores of the judges, our criterion will be that of statistical control. A state of statistical control is said to exist when the data fall within the control limits in a random manner.

Let us examine all the points on Charts I and II which do not appear to meet the criterion of statistical control. Comparing the control charts for $\overline{X}$ of all three tasters, it is obvious that all three average scores for the second week are lower than those for the first and the third weeks. Likewise, the average scores during the 5th and 6th weeks

are lower than for the week immediately preceding and following. All three series thus show the same pattern or trend for these periods and therefore lack randomness; they are not in a state of statistical control. In tracing down the cause of this lack of control, it might be found, for example, that there had been faulty preparation of the food in the 2nd week and a change of the judging time from morning to after lunch during the 5th and 6th weeks.

Since Chart I has a narrower spread between the control limits and hence is more sensitive than Chart II, it provides additional evidence for the lack of control during the above mentioned periods. In Chart I the average scores of tasters B and C for these periods fall outside the control limits; however, this does not hold true for taster A. His scores fall inside the control limits for the 2nd, 5th, and 6th weeks, and outside for all others—contrary to what might be expected. A study of his central line for averages in Chart II, which is higher than those of both tasters B and C by approximately one unit, reveals that taster A's scores were consistently too high. If his scores were lowered by one unit, then his points for the 2nd, 5th, and 6th weeks in Chart I would also fall outside the control limits. For these weeks, as was noted above, assignable extraneous causes had been detected.

In Chart I taster B's lack of control during the 3rd week, as indicated by both control charts for $\overline{X}$ and σ, could be attributed to reduced taste sensitivity because of a cold. But because no explanation could be found for the lack of control that most of taster C's scores showed in both Charts I and II, it might well be concluded that he was not sufficiently trained in the scoring of the food product in question.

Selection of a tasting panel. The first requirement for an efficient taster is his reliability, that is, his ability to reproduce his scorings when tests are repeated a number of times under essentially the same conditions. The control charts for $\overline{X}$ and σ indicate this reliability by the narrowness of the spread between the control limits. In addition, they provide a means for recognizing and interpreting changes in a taster's behavior by analyzing continuously his average scores and standard deviations in order to see if they are in a state of control. However, reliability does not insure validity, that is, correspondence of the scoring to the pre-established standard. A taster's judgments, although reliable, might not be near enough to the pre-established standard and therefore might not be valid. If the control chart for $\overline{X}$ is constructed around the pre-established standard as central line it can demonstrate a taster's validity in scoring.

In terms of the control chart method, it can be said that in order to be a good taster, average scores must be statistically controlled around

the pre-established standard as the central line with a narrow spread between the control limits; likewise, the standard deviations must be in a state of statistical control.

In the hypothetical data discussed above, it would appear that judge C is a poor taster. A and B are good tasters; divergences from conditions of statistical control are due to clearly assignable causes. Therefore, the scores of tasters A and B can normally be pooled for effective tests of food quality.

Having selected "good" tasters, the scores of two or more of them can be pooled to secure the best quality test of the product in question. A good taster's score might be compared to the results of weighing on a good scale. A scale is exact if it is carefully gauged (validity); if on repeated occasions its weight measurements are nearly equal (reliability); and if the deviations in its weight measurements are no greater than can be attributed to chance (statistical control). Obviously, one should choose the best scale available and should weigh an object a second time on the same or an equally good scale to make sure that no uncontrolled factor influenced the first measurement. If there were no great discrepancy between the two results, the average of them would, in practice, be taken.

Similar reasoning applies to the pooling of the scores of the tasters. The score of one good taster could be taken and then checked against the score of a second good taster. The two scores could then be averaged. Although in certain instances a larger number might profitably be pooled, for practical purposes, an average of 2 or 3 good scores might well suffice. If it seems desirable to check the difference between the pooled scores continuously, a control chart might be constructed to show whether they lie within a narrow enough spread. Moreover, applying a test of significance will show whether or not the difference between the pooled scores is significant.

When training new tasters no scores should be pooled. Lack of control during this period, as evidenced either by an unusual pattern or by points falling outside the control limits, would indicate that further improvement in the taster's performance is necessary. Thus the length of the training period might be determined from the control chart. The decision as to whether or not a state of statistical control is attained will partly depend upon engineering judgment. "It usually will be found advantageous to have at least 25 subgroups for this purpose. Where speed is important, and it is desirable to get the control going with a minimum of delay, as few as 10 successive subgroups may be used initially."[9]

[9] See footnote 7 *supra*, p. 216.

ADDITIONAL USES OF THE CONTROL CHART METHOD IN FOOD TESTING

The above discussion is based upon the assumption that the variability of the scorings can be assigned to the behavior of the judges. However, assignable causes for the lack of control encountered in organoleptic testing may also be sought in certain technical factors of the experimental process that may affect quality characteristics, such as methods of processing and packaging or the source of supply of the material. In this case control charts can be used in grouping the data on the basis of these technical factors.

It might be worthwhile to note that instead of constructing control charts for each individual taster separately on the basis of time, as was done above, the behavior of the tasters could also have been analyzed by one single control chart using each taster's scores as a subgroup. This method would provide an overall comparison of the tasters' efficiency but would not register temporal changes in their behavior.

In the marketing of food, the quality of the product often is tested by official graders. These graders may test the food either organoleptically or with mechanical or other devices applying previously established standards. Their test results may be subjected to an analysis by the control chart method similar to that discussed in this paper.

When establishing a standard for taste scoring, this standard should, whenever possible, take into account results of consumer-preference tests. However, no such authority is available if the tasters are to score a newly manufactured food product which has not yet been submitted to the consumer for his preference. In this case, a standard may be established by a panel of experienced tasters whose previous scoring on other foodstuffs has been found by the control chart method to be satisfactory.

In taste scoring according to a standard based solely on the results of consumer-preference tests, it is well to correlate the scores with results from laboratory tests. A careful interpretation of this relationship, taking into account the changing nature of human desires and needs, will be important for attaining effective quality control.[10]

SUMMARY

From the foregoing analysis, it can be seen that the control chart method applied to organoleptic testing is useful in the following ways: (1) it can assist in the selection of a good tasting panel, that is, persons

[10] A. G. Ashcroft, "The Interpretation of Laboratory Tests as Quality Indices in Textiles," *American Dyestuff Reporter*, Vol. 33, No. 24, November 20, 1944.

whose tasting scores are valid and meet the requirements of statistical control; (2) it determines what specific tasting scores must be examined to find out whether assignable extraneous causes are present; (3) it provides a means of minimizing the losses arising from two sources of error: (a) failure to pool the results of those tasters who may be pooled as checks against each other; and (b) pooling results that should not be pooled.

The control chart method is advantageous during an initial training period of judges, since its records indicate the length of time necessary for a required degree of control in scoring. A sufficient reduction in the spread of the control limits on $\overline{X}$ in subsequent scoring would indicate that better standards of performance can be set for the immediate future.

In general the control chart method can be useful, not only in testing foods organoleptically, but also in grading the quality of food products objectively according to a fixed standard.