

RSS-30

DEVELOPMENT OF A SCALE FOR MEASURING SOLDIERS'
FOOD PREFERENCES^{a, b}

LYLE V. JONES

University of Chicago, Chicago, Illinois

DAVID R. PERYAM

Quartermaster Food and Container Institute for the Armed Forces, Chicago, Illinois

AND

L. L. THURSTONE

University of North Carolina, Chapel Hill, N. C.

(Manuscript received May 21, 1955)

Acceptability, always an important consideration in food development and utilization, is particularly so in problems of mass feeding such as those encountered in designing rations for the Armed Forces. Military rations must be adjusted to the preferences of the entire population of Service men. Even foods that are extremely well-liked, but by only a small proportion of the consumers, are unsuited for military use. Items must be selected which have satisfactory average preference and are disliked by as small a proportion of the population as possible. For efficient selection, methods are required which will determine acceptability in different kinds of situations, including laboratory and field pretests of actual food items, and which will serve for investigations of food preferences. All such tests depend on the use of psychological measurement to reduce to a common scale the subjective attitudes of many people. Experience has shown that the approach commonly called the rating scale method, or, more completely, the method of successive intervals, is the most appropriate and efficient.

In 1949 a device known as the "hedonic scale" was developed at the Quartermaster Food and Container Institute for the Armed Forces and has become the standard instrument for use by the QM Corps in laboratory and field tests of acceptability (6). Although it has provided usable information about food preference, certain deficiencies in the scale were noted. Since accurate measurement of food preferences is vital in food research, it became important not merely to correct recognized defects, but to establish with a reasonable degree of certainty a method which would be optimal for military use. In 1951 the Psychometric Laboratory at the University of Chicago undertook such a project.

^aThis paper reports research undertaken in cooperation with the Quartermaster Food and Container Institute for the Armed Forces, Chicago, Ill., and has been assigned number 524 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 Defense.

^bPresented at the Fourteenth Annual Meeting of the Institute of Food Technologists, Los Angeles, California, June 29, 1954.

[1]

TECHNICAL LIBRARY
U. S. ARMY
NATICK LABORATORIES
NATICK, MASS.

SCALE FOR MEASURING FOOD PREFERENCES

Problem. Inspection of a rating scale may suggest that widths of the intervals should be equal. However, there is never assurance that any one interval is of the same width, psychologically, as any other. In fact, typically, there is evidence of gross inequality. The reasonable objective is a rating scale for which no one would question that the successive intervals are in the proper ordinal position, and where all subjects understand and use the intervals in about the same way. When that has been achieved, the variance in the ratings of a particular food may be interpreted as indicating different levels of preference for that food, rather than different ways of understanding the rating scale.

The choice of words or phrases to label the scale intervals is of first importance, since these verbal anchors serve both to convey the idea of the successive order of the intervals and to make clear to the respondents the meaning of the response continuum. The value of a scale will be reduced to the extent to which the words and phrases are ambiguous, or are not definitely in an order of meaning corresponding with the physical order of the scale intervals. Scales may vary in other ways, too. Among the most important are (a) the number of intervals, (b) whether or not the scale is balanced, i.e., has an equal number of positive and negative intervals, and (c) whether or not a "neutral" category is included. All of these variables are included in the present study.

PROCEDURE AND RESULTS

The research reported here involves a number of interrelated phases. The first task was to develop and evaluate a potential "food preference vocabulary." Following this, two series of scales, each of which embodies certain hypotheses regarding the other important variables, were designed and evaluated in field surveys. Pertinent procedural details are included in the discussion of results.

Selection of descriptive phrases. Fifty-one words and phrases were selected for investigation. Part of this list resulted from a pilot study with a group of Army men; other elements were included because of their frequent use in preference questionnaires or their apparent logical suitability. Subjects were approximately 900 soldiers from Fort Lee, Virginia, selected on the basis of educational background to be representative of Army enlisted men. Figure 1 shows the rating scale and examples of items that were included. The subjects were told: "The items are words and phrases that people use to show like or dislike for foods. For each item make a check mark in the box which best shows what the word or phrase means to you."

The methods of analysis used in this phase of study have been described elsewhere (4). Briefly, the analysis provides for determination of a psychological continuum of meaning which exhibits the characteristics of an equal interval scale, the method being based on the assumption that each phrase has a modal meaning about which the various meanings attributed to it by the respondents are normally distributed. A scale value and standard deviation are derived graphically for each item. The former may be considered the "average meaning" for the phrase, and the latter a measure of its relative ambiguity. In Table 1 appear these indices for all of the phrases. It will be noted that size of the standard deviation is not independent of scale value, for as scale values depart from zero, standard deviation values tend to increase. This result correctly is interpreted as indicating relatively greater ambiguity of meaning of "extreme" phrases than of "neutral" phrases. That this is a reasonable finding is clear; as the meaning of a phrase departs from that of neutrality, it becomes more likely that individuals will exhibit greater disagreement as to the precise position of the phrase on the meaning continuum.

An important aspect of the distributions, which is not apparent from the numerical data alone, is shown by graphical plots (on binomial probability paper) of cumulative proportions of responses against the scale values of the boundaries of the successive

	GREATEST DISLIKE		NEITHER LIKE NOR DISLIKE					GREATEST LIKE	
	-4	-3	-2	-1	0	+1	+2	+3	+4
1. LIKE INTENSELY									
2. DISLIKE SLIGHTLY									
3. STRONGLY DISLIKE									
4. DESPISE									
5. MILDLY LIKE									
6. WELCOME									
7. NOT PLEASING									
8. PREFERRED									
9. DISLIKE VERY MUCH									
10. LIKE NOT SO WELL									

Figure 1. Scale and examples of phrases used in the meaning study.

intervals. Departures from linearity on these graphs illustrate failures in the assumption of normality. Graphs for three of the phrases are shown in Figure 2. Included are two of the six phrases which show marked departure from normality, together with one phrase, "preferred," for which departure is slight. "Dislike moderately" illustrates a positively skewed distribution. A significant number of men marked it on the "like" side of neutral. "Average" exhibits a bimodal distribution; one group of men marked it at the center and another group placed it two steps above the center. Each instance of a non-normal distribution can be diagnosed as an indica-

SCALE FOR MEASURING FOOD PREFERENCES

TABLE 1
Scale values and standard deviations for 51 descriptive phrases included in
the word meaning study

Phrase	Scale value	Standard deviation
Best of all	6.15	2.48
Favorite.....	4.68	2.18
Like extremely.....	4.16	1.62
Like intensely.....	4.05	1.59
Excellent.....	3.71	1.01
Wonderful.....	3.51	.97
Strongly like.....	2.96	.69
Like very much.....	2.91	.60
Mighty fine.....	2.88	.67
Especially good.....	2.86	.82
Highly favorable.....	2.81	.66
Like very well.....	2.60	.78
Very good.....	2.56	.87
Like quite a bit.....	2.32	.52
Enjoy.....	2.21	.86
Preferred.....	1.98	1.17
Good.....	1.91	.76
Welcome.....	1.77	1.18
Tasty.....	1.76	.92
Pleasing.....	1.58	.65
Like fairly well.....	1.51	.59
Like.....	1.35	.77
Like moderately.....	1.12	.61
OK.....	.87	1.24
Average.....	.86	1.08
Mildly like.....	.85	.47
Fair.....	.78	.85
Acceptable.....	.73	.66
Only fair.....	.71	.64
Like slightly.....	.69	.32
Neutral.....	.02	.18
Like not so well.....	-.30	1.07
Like not so much.....	-.41	.94
Dislike slightly.....	-.59	.27
Mildly dislike.....	-.74	.35
Not pleasing.....	-.83	.67
Don't care for it.....	-1.10	.84
Dislike moderately.....	-1.20	.41
Poor.....	-1.55	.87
Dislike.....	-1.58	.94
Don't like.....	-1.81	.97
Bad.....	-2.02	.80
Highly unfavorable.....	-2.16	1.37
Strongly dislike.....	-2.37	.53
Dislike very much.....	-2.49	.64
Very bad.....	-2.53	.64
Terrible.....	-3.09	.98
Dislike intensely.....	-3.33	1.39
Loathe.....	-3.76	3.54
Dislike extremely.....	-4.32	1.86
Despise.....	-6.44	3.62

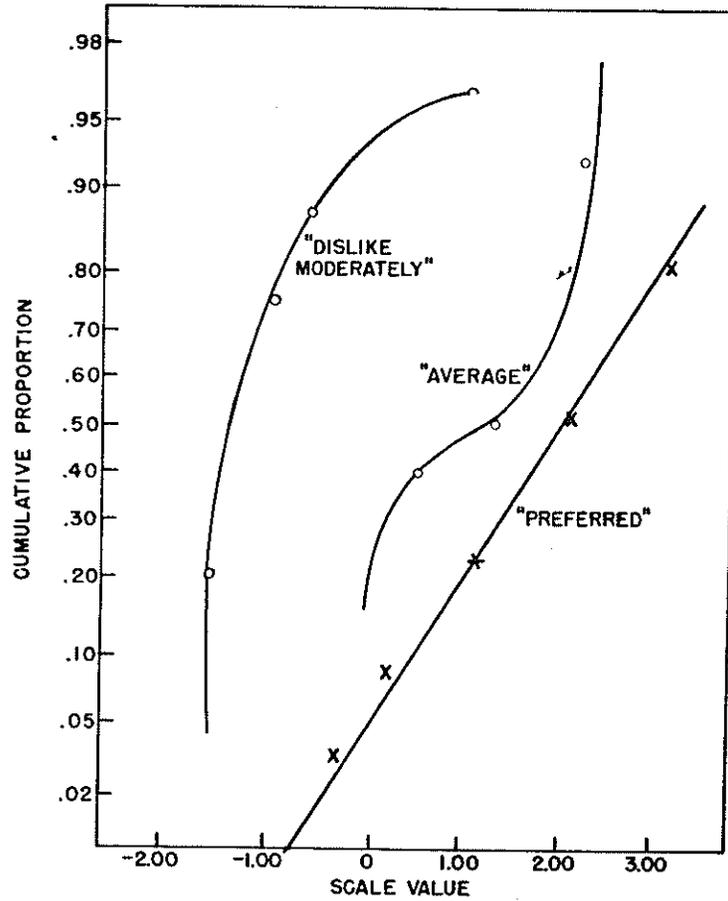


Figure 2. Graphical plots for three descriptive phrases displaying different types of distribution.

tion of some particular confusion regarding the meaning of the phrase.

On the basis of these findings, descriptive phrases could be selected for use in preference scales on the basis of their known "average meanings," low ambiguity, and slight departures from normality.

Comparison of scales. The nine different scale types shown in Figure 3 were investigated. Note that they vary in length and that various combinations of phrases are employed to describe the intervals. The middle interval is eliminated in Nos. 4, 7, and 9. Nos. 8 and 9 are "unbalanced," with fewer "dislike" than "like" categories. No. 1 is the hedonic scale currently used by the QM Corps. Scales 1-5 were included in the first field test, conducted in June 1952. The respondents were 3600 enlisted men sampled from the four Army posts on the eastern seaboard. The second field test was administered at Fort Bragg, N. C. in August 1953 to 1800 men, and included scales 1 and 6-9.

Each test was ostensibly a preference survey of 20 food items which had been selected, on the basis of previous survey results, to cover a wide range of preference. All 9 questionnaires studied include the same food items, and differ only in regard to the rating scales. Respondents were simply instructed to check the reply which best showed how much they liked or disliked each food. Questionnaires were administered in class sessions, each of which included no more than 100 men, and the five scale types were systematically and evenly distributed in each group.

SCALE FOR MEASURING FOOD PREFERENCES

SCALE NUMBER	NUMBER OF INTERVALS	PHRASES DEFINING SUCCESSIVE INTERVALS												
		DISLIKE EXTREMELY	DISLIKE VERY MUCH	DISLIKE MODERATELY	DISLIKE SLIGHTLY	NEITHER LIKE NOR DISLIKE	LIKE SLIGHTLY	LIKE MODERATELY	LIKE VERY MUCH	LIKE EXTREMELY				
1	9	DISLIKE EXTREMELY	DISLIKE VERY MUCH	DISLIKE MODERATELY	DISLIKE SLIGHTLY	NEITHER LIKE NOR DISLIKE	LIKE SLIGHTLY	LIKE MODERATELY	LIKE VERY MUCH	LIKE EXTREMELY				
2	9	DISLIKE EXTREMELY	DISLIKE VERY MUCH	DISLIKE	MILDLY DISLIKE	NEUTRAL	MILDLY LIKE	LIKE	LIKE VERY MUCH	LIKE EXTREMELY				
3	7	DISLIKE EXTREMELY	DISLIKE VERY MUCH	DISLIKE VERY MUCH	MILDLY DISLIKE	NEUTRAL	MILDLY LIKE	LIKE VERY MUCH	LIKE EXTREMELY					
4	6	DISLIKE EXTREMELY	DISLIKE VERY MUCH	DISLIKE VERY MUCH	MILDLY DISLIKE	MILDLY DISLIKE	MILDLY LIKE	LIKE VERY MUCH	LIKE EXTREMELY					
5	5	DISLIKE EXTREMELY	DISLIKE VERY MUCH	DISLIKE EXTREMELY	DISLIKE	NEUTRAL	LIKE	LIKE EXTREMELY						
6	9	DISLIKE EXTREMELY	DISLIKE VERY MUCH	DISLIKE FAIRLY MUCH	DISLIKE SLIGHTLY	NEITHER LIKE NOR DISLIKE	LIKE SLIGHTLY	LIKE MODERATELY	LIKE VERY MUCH	LIKE EXTREMELY				
7	8	DISLIKE EXTREMELY	DISLIKE VERY MUCH	DISLIKE MODERATELY	DISLIKE SLIGHTLY	DISLIKE SLIGHTLY	LIKE SLIGHTLY	LIKE MODERATELY	LIKE VERY MUCH	LIKE EXTREMELY				
8	8	DISLIKE EXTREMELY	STRONGLY DISLIKE	MILDLY DISLIKE	NEITHER LIKE NOR DISLIKE	NEITHER LIKE NOR DISLIKE	LIKE SLIGHTLY	LIKE MODERATELY	LIKE VERY MUCH	LIKE EXTREMELY				
9	7	DISLIKE EXTREMELY	STRONGLY DISLIKE	MILDLY DISLIKE	MILDLY DISLIKE	MILDLY LIKE	LIKE FAIRLY WELL	LIKE QUITE A BIT	LIKE VERY MUCH	LIKE EXTREMELY				

Figure 3. Scales investigated in the field surveys.

The following criteria were established to determine the relative adequacy of the scales: (a) ease of completion as shown by the amount of time required, (b) reliability as shown by the accuracy with which respondents duplicate results on an alternate form re-test, and (c) the amount of information obtained about the relative preference values of the group of foods.

(a) Time required for completion.

If there were major differences in the time required to complete the questionnaires, this would be an important criterion of relative efficiency of measurement. Proctors

recorded the time required by each respondent to complete the questionnaire in the 1952 survey, which included scales 1-5. These scales vary in length from 5 to 9 intervals. As expected, completion time is found to increase with the number of intervals; however, the difference between the shortest and longest median times is only 14 seconds. Obviously, this criterion needed no further consideration.

(b) *Reliability.*

Approximately 1250 subjects in the 1952 survey and all 1800 men in the 1953 survey were retested with a second questionnaire of the same scale type as the first. The retest took place as soon as all subjects had finished the first questionnaire. The same 20 food items appear on the second form, but are arranged in a different order. Product-moment correlations between responses to the same food items on these alternate forms are used to assess the reliabilities of the scale.

One striking result is the finding that reliability for certain food items is consistently high, whereas for others it is consistently lower. Differences among the reliability coefficients for various food items are much greater than differences among scales. To cite the extreme examples, the average correlation is +.92 for iced coffee, but only +.70 for jellied fruit salad.

TABLE 2
Averages¹ of test-retest reliability over 20 food items for nine scale types

Scale number	Number of intervals	Characteristics	First survey (1952)	Second survey (1953)
1	9	balanced, neutral	.821	.836
2	9	balanced, neutral	.833	
3	7	balanced, neutral	.848	
4	6	balanced, no neutral	.819	
5	5	balanced, neutral	.824	
6	9	balanced, neutral		.857
7	8	balanced, no neutral		.826
8	8	unbalanced, neutral		.814
9	7	unbalanced, no neutral		.826

¹ After transformation to Fisher's *z* statistic (2).

Table 2 gives the reliabilities for the 9 scale types, obtained by averaging over all food items. They cover the restricted range from +.814 to +.857, and show no consistent relationship with the number of intervals on the scale. The differences among reliabilities of the scales are at a level for which statistical significance is doubtful; and certainly they are of little practical importance.

(c) *Transmitted information.*

The most meaningful criterion for assessing the relative values of scales is the amount of information transmitted (3, 5). High transmitted information values indicate discriminating responses to the food items included in the survey, i.e., distinct and different distributions of responses for the various foods with a high level of agreement among the ratings for each. Since the ultimate objective is to have a scale as sensitive as possible to all differences among food preferences, the amount of transmitted information takes on great importance.

If other factors are held constant, the potential amount of information increases with the number of intervals. This follows from the nature of the information index, since, with more response intervals, there is greater opportunity for fine discriminations among stimuli. Several empirical studies have confirmed this relationship (1), and it is borne out by results of the first survey. With one exception the information values increase as the number of intervals increases from five to nine. The 6-interval scale, No. 4, is an exception which led to the hypothesis that elimination of the mid-point, or neutral category, would increase the transmitted information. Results of the second survey tend to confirm this hypothesis.

Figure 4 is a graph of the transmitted information values obtained in both surveys, in which the scales are grouped according to number of intervals. Higher information values tend to go with the longer scale; however, the values associated with the two

SCALE FOR MEASURING FOOD PREFERENCES

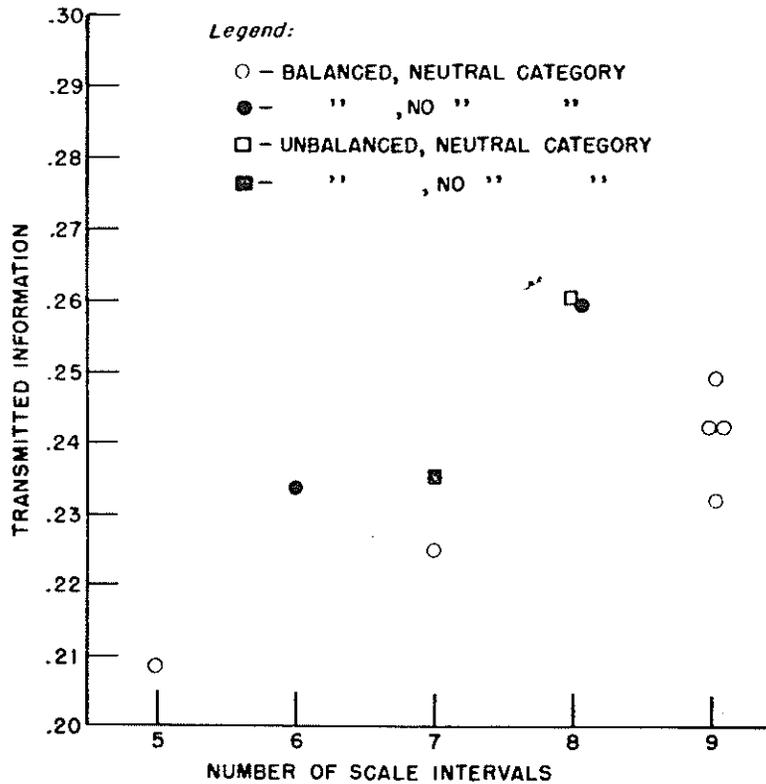


Figure 4. Transmitted information in relation to number of scale intervals.

8-interval scales are 5-10% higher than those for the 9-interval ones. This advantage appears both when the center category is omitted (No. 7), and when the *dislike moderately* category is omitted, leaving an unbalanced scale (No. 8). Even though the numerical differences among these indices are small, the consistency with which they appear is noteworthy. An appropriate non-parametric statistical test leads to rejection, at the .01 significance level, of the hypothesis that the differences are independent of number of scale intervals.

DISCUSSION

In one sense this research has failed to attain its objective, since no uniquely superior scale has as yet emerged. The similarities among the scale types investigated, particularly with regard to ease of completion and reliability, are more striking than the differences. The differences in transmitted information, although significant, are numerically small. However, the conclusion that it makes but little difference how a scale is constructed does not follow, because the range of scale types investigated was highly restricted. Selection and placement of the descriptive phrases on the basis of the vocabulary study was undoubtedly a most important factor; also, only those scale lengths were included which previous work has shown to be near the optimal range. No "poor" scales were specifically included as controls and the hedonic scale previously developed at the QM Food and Container Institute for the Armed Forces, Chicago, (No. 1) happened

to have many of the characteristics later shown to be desirable; thus the data give nothing like a "best-worst" comparison.

More often than not rating scales used for measuring preference and various qualities of foods have been balanced, with an equal number of positive and negative intervals, and have included a neutral point. Apparently this has been due to logical considerations, rather than experimental evidence. The present studies failed to find any evidence that either characteristic is advantageous. The two 8-interval scales, one balanced and the other unbalanced, gave almost identical information values; with the two 7-interval scales, the trend is in favor of the unbalanced scale. The neutral category was omitted in an 8-interval, a 7-interval, and in the 6-interval scale. Again, this omission caused no loss of information, but rather tended to increase transmitted information.

CONCLUSIONS

These results have implications for the practical problem of evaluating foods in terms of human preferences as well as for psychological measurement theory. Conclusions believed most pertinent to the food technologist are as follows:

- a. Descriptive phrases may differ greatly in ambiguity.
- b. They differ also in the level of preference implied, and this cannot always be predicted on an *a priori* basis.
- c. Increasing the length of a scale, up to nine intervals, is related to only a negligible increase in the time required for completion.
- d. Test-retest reliability, within the range of five to nine intervals, is relatively invariant.
- e. Longer scales, up to nine intervals, tend to be more sensitive to differences among foods.
- f. Elimination of the "neutral" category seems to be beneficial.
- g. Balance, i.e., an equal number of positive and negative intervals, is not an essential feature of a rating scale.

LITERATURE CITED

1. BENDIG, A. W., AND HUGHES, J. B. II. Effect of amount of verbal anchoring and number of rating-scale categories upon transmitted information. *J. Exptl. Psychol.*, 46, 87-90 (1953).
2. FISHER, R. A., AND YATES, F. *Statistical Tables for Biological, Agricultural and Medical Research*. Hafner Publ. Co., Inc., New York: 1953.
3. GARNER, W. R., AND HAKE, H. W. The amount of information in absolute judgments. *Psychol. Rev.*, 58, 446-459 (1951).
4. JONES, L. V., AND THURSTONE, L. L. The psychophysics of semantics. *J. Applied Psychol.*, 39, 31-36 (1955).
5. MILLER, G. A. What is information measurement? *Am. Psychologist*, 8, 3-11 (1953).
6. PERYAM, D. R., AND GIRARDOT, N. F. Advanced taste test method. *Food Eng.*, (July, 1952).