Peryam & Kroll Research Corporation

The 9-Point Hedonic Scale

Dr. David R. Peryam's early papers on the most widely used sensory scale in the world

P&K research

THE 9-POINT HEDONIC SCALE

Dr. David R. Peryam's early papers on the most widely used sensory scale in the world

Introduction by Dr. Edgar Chambers, IV

Peryam & Kroll Research Corporation

Dedication

Dave passed away more than 5 years ago, but he has never been far from our minds. Not only is he remembered in the name of the company we started almost fifty years ago, not only do we often wish he were with us to offer his experience and wise perspective, not only do we still use some of the procedures and systems he helped establish for our sensory and marketing research work, but on a daily basis we are engaged in a science founded on Dave's early work.

Since his death we have tried to honor his memory in a variety of ways -- from an ASTM Award in his name and memorial services to sponsoring student activities and internship programs. However, nothing could ever be a better memorial to Dave than his own achievements.

What better way, then, to recognize those achievements than to guarantee that his work be easily available to students, historians and sensory professionals. Hence, this publication, which focuses on the 9-Point Hedonic Scale, developed by Dave and today the most widely used sensory scale in the world.

The Scale is so elegant the theory, proofs and protocols are encapsulated in one paper, although we are printing the four papers that, to our knowledge, comprise everything Dave published related to the Scale per se. The fifth paper in this publication is a sample of the studies Dave did throughout his life to continuously confirm the contemporary value of the Scale. A personal letter on the Scale's use as a category appraiser will, I hope deepen your appreciation of Dave's integrity (and his humor).

I recommend that every student and professional read these papers closely. They present a picture of a brilliant and modest scientist -- pointing out what he felt were shortcomings in the Scale while demonstrating its reliability. They also offer a look into sensory research that should be a renewal for professionals, and an inspiration for students. Dave always had a way of laying open basic truths and calling our attention to what was really important.

Beverley J. Kroll
May, 1998

THE 9-POINT HEDONIC SCALE

David R. Peryam, Ph.D. 1915 - 1992
"The Father of Sensory Science"

Table of Contents

Introduction

There are few sensory techniques as important and none that are more important than the hedonic scale. The papers presented in this booklet are some of those that David Peryam wrote or helped write that discuss the "development" and use of the hedonic scale.

Each time I read these and other related papers I grow more amazed at the depth and breadth of the research that was conducted on the development and application of a "scale for measuring ...food preferences." Of course we now refer to the phenomena measured by the hedonic scale as "liking" rather than "preference," but the development of the method remains seminal to work in sensory analysis and has been used by marketing research and vast numbers of product researchers in foods and other product areas.

The papers presented give a picture of the development and use of the scale and the vast amount of research that was conducted, first with soldiers and later with a more generalized consumer population. I hope as people read these papers and others that David authored or co-authored on the hedonic scale that the thoroughness of the research on the scale will astonish people and remind us that it is not just products that need to be evaluated, but that the methods we develop and select to help us assess the products also must have a sound pedagogical and empirical basis. I hope that people will be a little "shocked" in reading these papers to find that the authors do not presuppose that there is only one right scale or one right way to do things and readily point out that "no uniquely superior scale has yet emerged." We should see from these papers that it is unacceptable simply to "believe" that a method works or doesn't work and we must constantly strive to evaluate and understand the truth about methods and their variations rather than simply accept unsubstantiated opinions.

The number and variety of food products used to "test" the hedonic scale is unparalleled in the sensory methods literature. How fortunate we all are that the development and testing of the scale was conducted mostly with governmental resources, which not only allowed, but encouraged, publication of the data. Such publication gives us a rare glimpse at many of the steps necessary to evaluate a sensory method. Regardless of the type of test, we must assess the impact of the language used, we must study how the validity of the method stands up across product categories, we must consider how the test can and will be implemented, and we must consider how modifications to the test affect the information we gain. All of those issues were examined in the early years of the "hedonic scale" and they continue to be examined today as the scale is applied to global populations of all ages and new problems of product understanding.

Edgar Chambers IV
May, 1998

Problem of Preference Gets QM Focus

DAVID R. PERYAM
QM Food & Container Institute
For the Armed Forces, Chicago

Also of prime importance is panel-establishment of quality... industry's aims are closely parallel in pursuing both these key questions

PLURALITY of test booth permits multiple evaluations of the samples.

PANEL of women goes into Action as QM studies food likes and dislikes. Feminine factor is specially pertinent in comparible checks on civilian products.

Several areas of investigation are involved in the determination of the acceptance of foods and description of foods in terms of their sensory properties. And in each of these areas there lie a number of specific problems.

Some of these problems relate to the mechanics of measurement, such as test design, questionnaire construction, and selection of observers. Others are of a research nature, such as investigation of the monotony factor and the gaining of knowledge of the elements in the acceptance equation.

Of the many specific questions encountered in food acceptance, two are of particular importance both to the armed forces and to industry. They pertain to -

- Preference evaluation of food, and
- Establishment of food quality by panel procedures.

Preference - or "degree of liking" - for a food, is one of the most important factors in acceptance. In fact, preference and acceptance are considered almost identical in the thinking of many people and in their experimental approaches to the acceptance problem. An influential factor is that preference is relatively easy to investigate. Also, there is justification for this attitude in that this factor appears to be the most constant single determinant of acceptance. It can be measured in the laboratory with fair assurance that the same direction and relative magnitude of effect will maintain in field tests or other practical conditions of use. Hence its measurement becomes a first order of business in any acceptance program.

Termed "Hedonic Value"

A discussion of meanings is in order here. Although "preference", in a strict sense, refers merely to choice, without regard to the reasons for that choice, by tacit agreement it has come to mean pleasantness or degree of liking. This apparently causes little or no confusion. Psychology has a specific name for this factor - "hedonic value." The name is a definite help in theoretical understanding, but it must be used with caution in the food research field, since it usually will require definition.

Several methods, all of them contributed by experimental psychology over 50 years ago, are available for measuring hedonic values. Assessment of comparative values, either between paired samples or by ranking in order, gives the "power-house" technics this purpose, because the results, within their limitations, are very specific. One limitation is that they provide only relative values. Another limitation, from a practical point of view, is the considerable amount of labor required per unit of information obtained. The rating-scale method offers better possibilities and is used far more often.

Rating scales for evaluating are in common use, but they present a confusing picture to the researcher. They vary in physical design, which may be important, and in the task that they set for the test subject, which is of vast importance. Frequently, a rating scale succeeds in doing no more than comparing similar items that are presented together, hence it has no advantage over direct-comparison methods, except that of saving labor. When the purpose of the testing is only selection of one item from among a number of similar alternatives, this is no reason for criticism.

Food research for the Armed Forces, however, faces a broader problem that may frequently confront industry also. Feeding in the armed services may utilize any common food. It may use common foods in somewhat unusual forms, or there may be developed new and unusual items. At the same time, military requirements, particularly for operational rations, are such that the number and variety of items to be included in a rations are severely limited. There is competition not only between variants of a single type (the typical industrial problem) but also between foods that are quite different from each other.

Since one basis of competition is acceptability, there is need of a technic which will measure the hedonic values of all manner of products on a common basis. A method is required that will, for example, do more than tell how one brand of pork and beans stands up to another—that is, also give some idea of the value of pork and beans in relation to canned biscuits, chocolate candy, or food bar "No. X-21."

The basis for the required method is found in the fact that people are seldom indifferent to the food they eat, but nearly always experience pleasure or displeasure in some degree. This response remains basically the same, varying only in direction and intensity, even though the foods which bring it about may be quite different from time to time. Therefore, this response can serve as the common basis for comparison. Moreover, the level of this feeling should have predictive value for actual acceptance.

9-Point Scale

A rating scale and a test procedure have been derived from this theoretical basis. The scale has nine points, and these points are given word descriptions ranging from "dislike extremely" to "like extremely". Further, the instructions are designed to direct the test subject's attention to his feeling about the food, rather than to the food itself. For this reason, "like" and "dislike" terms are used instead of the more technically correct "pleasant" and "unpleasant", and also because they are more likely to mean the same to all persons.

The length of the scale was determined experimentally. Replicate testing of items of varying hedonic value showed that responses were repeated more consistently when the scale had 9, rather than 5, 7, or 11 points. Also, the test as a whole discriminated between the items as well with the 9-point scale as with any of the others.

This scale is being used constantly for consumer preference tests in the laboratory and has also been employed in field testing. In general, the results have been good. The method has a satisfactory reliability when repeated on similar groups, although its validity as a measure of hedonic value, as such, can hardly be determined until there is available some accurate, unquestioned method of determining likes and dislikes.

Although laboratory results with this scale give considerable insight into what will happen in the field, there is plenty of evidence that it cannot in itself predict final acceptance. But this in no sense invalidates the method, since hedonic value is but one factor in acceptance.

The possibility of developing improved ways of interpreting results gives hope that the method may become even more valuable. Simplicity demands that, if possible, results be reduced to a single index. To do so at present, the scale points are numbered from 1 to 9, and the arithmetic mean of the points checked is used as the desired index.

However, with any group of consumers, the assigned ratings tend to spread over a wide range of the scale. This may be considered as a true reflection of the well-known fact that actual likes and dislikes for a particular food also vary widely. When large groups of observers are used, the frequency distributions of assigned values tend to fall into certain patterns, and it seems likely that these patterns may have a significance for prediction of hedonic value and acceptance far beyond what is possible with the mean rating.

Establishing Food “Quality”

The second methodological problem is that of establishing quality of a food by trained panel judgment of sensory properties. Such tests are probably used more often than any other type in sensory food evaluation throughout the country. “Quality” in most such testing is defined poorly, if at all. The ability to apprehend quality is supposed to come only through experience, in amount and nature also undefined.

The Institute’s acceptance program is committed to defining excellence in terms of the consumer’s response. Hence, attaining a judgment of “good quality” is not an end in itself. Interest lies in using the method to predict the “like-dislike” response of consumers. In this way “quality” serves as an intermediate criterion.

The method has value for two reasons: First, it usually gives better discrimination between items than does a consumer preference test. And, second, it is easier to conduct, since it requires fewer people and doesn’t bring up as serious a sampling problem.

Judgment of quality is also approached by the rating-scale method, using a scale in which the points are described as levels of quality. There may be some question as to whether a valid distinction can be made between this and the hedonic value method. It is quite possible that a naive observer would respond the same way on both. The difference lies in the kind of response which is asked for.

The hedonic value method is designed to call forth a naive, immediate response, while the other is gaged to encourage the exercise of judgment.

The distinction is made valid by using in the latter test only observers who have been selected and trained for ability to judge. Quality is defined for them as that combination of sensory properties which is most likely to please the consumer, and they are given as much information as is available about the trend of consumer preference in any particular case.

This approach is based upon the premise that observers can be trained to disregard their own individual preferences and to evaluate foods according to their knowledge of the behavior of others. Here, validity is an immediate and vital concern, since if this test cannot predict consumer preference with at least passable accuracy, its efficiency avails nothing. However, this can easily be checked in the laboratory.

Although at the present time the quality judgment method utilizes primarily a 9-point balanced scale similar to that employed for hedonic value, there is evidence that having a fixed scale type and scale length may be of no advantage, and that both reliability and discrimination might be increased by adjusting the scale to suit the product and the training of the observers. Nor is there anything to be gained here in having all ratings expressed in a single index, since it is hardly significant to compare the excellence of one type of product with that of another. The test is expected only to compare items of the same type by reference to the standard of quality for that type.

Trained Panels

A further problem has been suggested in discussion of both the judged quality method and the analytical use of sensory testing for research purposes. This is the matter of training. Any laboratory which does sensory testing must either consider this factor, or else rather pointedly ignore it. Great individual differences exist in sensitivity and skill. It is also evident that peoples' reactions tend to change when they are continually subjected to test situations. Any sensory testing program should have a theoretical approach and (if possible) a practical policy which take such conditions into account.

Although use of the small trained panel is almost traditional, often it is not clear just what the panel brings to bear, beyond a suggestion of an aura of some sort of specialization. It may be represented as having special technical knowledge or long experience. Too frequently, its only qualification is availability. The result has been that to dub a group a "trained panel" has had little meaning in itself.

In the Institute's acceptance program, "trained panel" is defined operationally by the process by which panels are established. Panels are set up for specific products or for particular types of investigation. A group of at least three times as many persons as are desired for the panel is tested for ability to detect typical differences that may be found in the product under investigation, and those persons showing obviously poor discrimination are discarded.

"JUST HOW SENSITIVE is he to odor?" That's one of the essential questions in rating a test subject for QM's special appetite studies. This Elsberg technic gives the answer.

Those remaining are intensively tested for ability to detect differences and are further evaluated as to stability of their judgments of quality upon replicate testing of a number of items. The resulting data permit ranking of all potential panel members according to skill and as many as are needed can be taken from the top ranks.

These panels might be considered more "selected" than "trained." However, the selection process itself provides a good deal of training. And after selection, the panel members are continuously provided with any information and instruction that will be helpful in orienting them.

Consumer-Preference Observers

Laboratory consumer preference testing, too, is never free from observer problems, and they are more serious than problems of trained panels since arbitrary solutions are less acceptable here. Sometimes there is a question of determining what type or class of people to use. But too often no selection is possible, and there is only concern over the validity of the test results. Seldom is a laboratory fortunate enough to have available a good sampling of the actual consumers of the product. The usual approach must be that of avoiding as many suspected biases as possible.

The Armed Forces acceptance program is no exception. Service men are the potential consumers, but are not available for laboratory preference testing. Those persons available to the Food Acceptance Laboratory of the Institute are 200 employees of the Institute, including technologists and administrative and office personnel, and about 500 persons from other activities of the Chicago Quartermaster Depot, including administrative and office personnel, as well as skilled and unskilled laborers. Candidates for the trained panels are drawn from this pool, and criteria have been set up for selecting persons from the same source for preference tests.

By policy, a consumer is any one who: (1) Has no technical knowledge of the product under investigation, (2) knows nothing about the immediate problem, and (3) has not been selected on a panel for the product concerned. Selection is random within these restrictions. This pool of observers probably represents a typical cross section of civilian food preferences, and the assumption must be made that the soldiers' preferences will not be found to differ markedly.

Where testing is done continually at one location, it is often suspected that a person's preferences will tend to change as he is subjected to the test situation again and again. Some laboratories even avoid preference testing, in the belief that this tendency completely nullifies results. Logically, it is quite possible that a person might cease responding with naive expression of likes and dislikes and, instead, express opinions derived from his testing experience. Yet, little is known about this change - even its existence is not proven.

If the effect is real, the point at which the change begins, its direction and amount, the course of its progress, and how to delay or prevent it, are matters of speculation. This represents a serious problem for the Institute, since the possibility of the existence of such an unknown in every result makes interpretation difficult.

Comparison With Industrial Problems

An approach to a solution is possible with the large pool of observers who are available. Most of the 200 personnel of the Food & Container Institute have participated in the testing program for several years and have had full opportunity to become "pseudo-trained," while most of the 500 remaining have had little or no experience. The first group will constitute a control against which to compare results from the larger group of naive observers, as its members acquire experience. This may give some insight into what, if anything, is happening.

It will have been noted that the Armed Forces acceptance problems are quite similar to those encountered in producing for the civilian market. However, two factors might be pointed out which complicate the usual problems for the armed forces, although they do not require any basic changes in approach:

First, food for Services' feeding must be designed not only for the conditions approximating the civilian norm, but also for use under conditions physical, physiological, and psychological extremes. Meals normally served at camps within the United States represent the former, while special purpose rations, such as for combat feeding, represent the latter. The influence of such extremes is an unknown in the acceptance equation and tends to make prediction even less exact.

Second, for reasons of economy and operational efficiency, a feeding program, or a ration, must be designed for an entire group. A new commercial product which pleases a small number - say 25 percent of the potential consumers - so that they buy it repeatedly, is considered a huge success. But a ration which pleases only 25 percent of the soldier-consumers is worthless to the military. A great many specialty items cannot be considered because of the necessity for pleasing the vast majority, and this need puts a premium on careful acceptance prediction for all rations.

But there is one condition which simplifies the Services' acceptance problem: Here, the consumer population is well-defined and is uniform in some of the most important characteristics. It is primarily male, young adult, and active. This will compensate in part for the difficulty represented by the other two factors.

Basic Research in Food Acceptance

At present, most food acceptance work is devoted to (1) problems of evaluating end-products and (2) to a type of service work for research in describing and controlling the sensory properties of foods. Scientific knowledge of acceptance seldom aids in original design of foods; this is still empirical, the realm of the expert and the artist. If, however, there are natural laws which govern acceptance and this seems certain—it should be possible not only to discover them, but also to use the knowledge to design acceptance characteristics into a food.

Present failure to achieve this end is due partly to lack of basic knowledge, also to poor organization and lack of confidence in what is known. These shortcomings create another area of interest for armed forces acceptance research. Investigations of the factors which affect acceptance, including those which pertain to the food itself, are underway at the Food & Container Institute and at outside laboratories under government research contracts.

Some of the major effects are known, such as climate, physical conditions under which food is eaten, physiological and mental states, and the whole gamut of food habits, training, and prejudices. But these effects have not been measured, their relative importance and their interactions have not been determined, nor do we know much about the variables which may operate within each major effect. Work completed and now in progress leads toward solution of some of these problems.

Gage of Appetite

The phenomenon of appetite has had considerable attention. The effects of various foods, exercise, vitamin deficiency, climate, emotional stress, and the administration of drugs have been observed. An important contribution has been determination of a relationship between taste and odor sensitivity and the momentary hunger-satiety state. It has been definitely established that most individuals are more sensitive when hungry and that subsequent loss of sensitivity depends upon the amount and kind of food eaten. The next objective in this line of research is to learn how to use this sensitivity index to evaluate foods.

The *modus operandi* of certain flavor agents is the subject of another phase of this basic work. It is usually assumed that all such agents are used to enhance acceptability, but it may not be known just how they bring this about, either the locus of the immediate sensory effects or how they are elaborated to positive acceptance behavior.

Monosodium glutamate is a good example. Work is underway to determine whether this substance actually stimulates all senses in the mouth, thereby causing a keener perception of flavor, as many of its proponents claim, or whether it is merely a blend of known basic tastes and adds a pleasant taste of its own, as does salt. In conjunction with this work goes the task of evaluating, by carefully controlled experiments, the effects of the substance upon consumer preference. An interesting finding in this area of flavor investigation has been that natural black pepper, in a great many of its normal usages, has no effect on acceptability of foods as measured by general consumer preference.

A-Ration Targeted

Food attitude, which is the resultant of a number of factors such as habit, training, social influences, personal opinion, and prejudice, is at present the subject of a series of studies that are of practical interest to the Food Service Division, and also are theoretically significant. The immediate objective is to obtain some knowledge of relative acceptability of the foods included in the master menu of the Army's A-ration.

A method has been established for this purpose. It utilizes a simple questionnaire in which the respondent indicates his degree of liking for no more than 45 foods at one time, the number being limited to avoid the dangers inherent in long questionnaires. Respondents are a probability sample of enlisted men drawn from the Army camps throughout the United States. A series of questionnaires are required to cover the 400-odd foods and recipes included in the master menu. Two segments of the study have been completed, and results have been satisfactory.

Not only will these studies enable the Food Service Division to alter the master menu toward economy and efficiency, but they will also be a fund of significant theoretical information. Later studies in the series will go beyond the basic "like-dislike" question.

Here are two paths research might profitably follow: take, first off, the elusive problem of monotony, where almost nothing is known beyond the simple fact that continued usage of some foods in certain situations tends to lower their acceptability.

Even to determine the true nature of this effect, to encompass the problem in a limited number of concrete propositions, thereby making it more accessible to research, might be no mean achievement. But more is needed. We must discover general principles governing this loss of acceptability, determine what proportion of it can be attributed to factors in the food and what proportion to training or other factors in the individual. Also we must devise means of anticipating the incidence and the extent of this negative influence.

It is apparent that a first approach can be study of the shifting patterns of attitude during development of monotony. Beyond that, the research area presents a large question mark.

The second problem concerns rations to be provided for survival in emergency situations. This is almost exclusively a military matter. What sort of acceptance characteristics are desirable in such rations? Should they be made pleasant, in order to boost morale and to insure full utilization? Or should they purposely be made drab or even unpleasant to discourage their use until hunger creates extreme motivation? Perhaps acceptance should not be considered at all, and they should be designed wholly for physiological value.

This whole range of theorizing is carried over into the design of the various survival rations now in use. The great obstacle to progress has been the difficulty of obtaining reliable information regarding human behavior toward food under physiologically and emotionally extreme conditions such as are expected in survival situations.

It is impossible to get accurate field reports. Most laboratory work has been done on animals. Hence, educated guesswork must stand in lieu of actual information. Again, here is a research problem where the field is wide open.

QM Pins Food "Likes" and "Dislikes" With

Advanced Taste-Test Method

David R. Peryam

and

Norman F. Girardot

Quartermaster Food and Container Institute for the
Armed Forces, Chicago

Small differences in similar foods, gross differences in checking general overall preferences, and group attitudes towards foods are now being quantitatively pegged using this hedonic scale adaptation

The above article is taken from the paper, "The Hedonic Scale Method of Measuring Food Preferences," designated as No. 385 in the series reporting research of the Quartermaster Food & Container Institute for the Armed Forces. The Institute points out that the views presented are those of the authors and are not to be construed as necessarily reflecting endorsement of the Department of Defense.

Approach to the Elusive

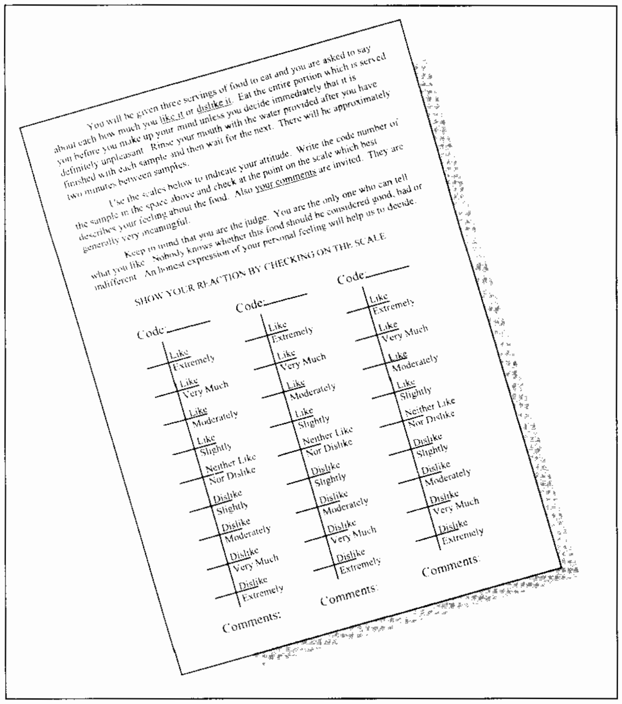
Attainment of reliability in consumer-preference evaluations of foods continues to be a troublesome problem, despite the various attempts at solution.

First, there has been general failure to achieve standardization.

Second, this failure has fettered development of confidence on the part of potential testers.

With a practical answer still elusive, the immediate logic is to entertain some partial solution that shows promise of lending stability.

Accordingly, the hedonic scale system is here advanced as a technique which may very well prove to supply this needed foundation for development of tomorrow's consumer preference methods. — *The Authors*



You will be given three servings of food to eat and you are asked to say about each how much you like it or dislike it. Eat the entire portion which is served you before you make up your mind unless you decide immediately that it is definitely unpleasant. Rinse your mouth with the water provided after you have finished with each sample and then wait for the next. There will be approximately two minutes between samples.

Use the scales below to indicate your attitude. Write the code number of the sample in the space above and check at the point on the scale which best describes your feeling about the food. Also your comments are invited. They are generally very meaningful.

Keep in mind that you are the judge. You are the only one who can tell what you like. Nobody knows whether this food should be considered good, bad or indifferent. An honest expression of your personal feeling will help us to decide.

SHOW YOUR REACTION BY CHECKING ON THE SCALE

Code: ______	Code: ______	Code: ______
Like Extremely	Like Extremely	Like Extremely
Like Very Much	Like Very Much	Like Very Much
Like Moderately	Like Moderately	Like Moderately
Like Slightly	Like Slightly	Like Slightly
Neither Like Nor Dislike	Neither Like Nor Dislike	Neither Like Nor Dislike
Dislike Slightly	Dislike Slightly	Dislike Slightly
Dislike Moderately	Dislike Moderately	Dislike Moderately
Dislike Very Much	Dislike Very Much	Dislike Very Much
Dislike Extremely	Dislike Extremely	Dislike Extremely
Comments:	Comments:	Comments:

Questionnaire used in laboratory to judge specific food preferences with the hedonic scale method. Tester checks point which best describes his reaction to food,

A technique has been developed at the *Quartermaster* Food & Container Institute that, we feel, offers notable progress in evaluating consumer preference of foods. It has been labeled the "hedonic scale method."

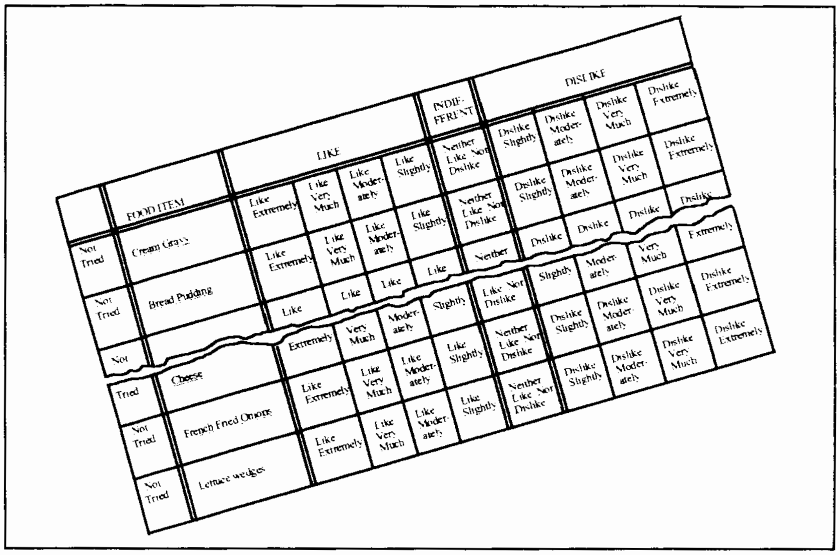
This basic approach is not new, since the method uses a variant of the well-known rating scale, introducing the hedonic value concept, which refers to the psychological range of "unpleasant" or "dislike" at the lower end to "pleasant" or "like" at the upper end.

Also, the problem of quality control of food flavors is completely divorced from this technique - thus keeping the analysis of consumer preferences on a separate plane.

This type of scale was first tried at the Institute in a comparison of methods of predicting soldiers' food choices. Results were encouraging, and in 1949 its suitability for the study of relatively permanent preference attitudes toward food was demonstrated. It was also shown to be adequate for laboratory use in measuring the response to foods as eaten. Forms and procedures were developed for both situations, and since then both kinds of applications have been used constantly.

Soldiers test coffee in mess hall at Fort Bragg during a recent field test conducted by QM Board. Immediately after finishing, men rate each sample on hedonic scale questionnaire. (Department of Defense photo)

Field Form Names Foods, Can be Used Any Time



	FOOD ITEM	LIKE				INDIFFERENT	DISLIKE			
Not Tried	Cream Gravy	Like Extremely	Like Very Much	Like Moderately	Like Slightly	Neither Like Nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely
Not Tried	Bread Pudding	Like Extremely	Like Very Much	Like Moderately	Like Slightly	Neither Like Nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely
Not		Like	Like	Like	Like	Neither	Dislike	Dislike	Dislike	Dislike
Tried	Cheese	Extremely	Very Much	Moderately	Slightly	Like Nor Dislike	Slightly	Moderately	Very Much	Extremely
Not Tried	French Fried Onions	Like Extremely	Like Very Much	Like Moderately	Like Slightly	Neither Like Nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely
Not Tried	Lettuce wedges	Like Extremely	Like Very Much	Like Moderately	Like Slightly	Neither Like Nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely

In field, preferences are registered on forms like this one - usually with nine items to be reviewed at one time. This type check-sheet was designed for studying soldier's preferences whether a few minutes or several days after eating foods under question. In scoring, each scale point is given a number.

The hedonic scale method is not considered a polished system, because pertinent questions as to its interpretation, its reliability, and the extent of its usefulness are as yet unanswered. Unquestionably, it can be improved. But it is described here in its present form because of the interest that it has aroused among many people concerned with preference evaluation of foods.

To present the hedonic scale method in terms of a set of rigid specifications would misrepresent the situation - since questionnaire forms can vary, and considerable latitude is allowable in the test. Forms and procedures now in use at the Institute will serve as a basis for this description with the critical features being emphasized and the points of permissible variation indicated.

In the standard questionnaire form used for laboratory consumer preference evaluations, two main parts may be recognized—the instructions and the scales. Instructions are generalized without reference to any particular food, and with provision for a maximum of three test items.

When the questionnaire is designed to measure general attitudes toward foods, the scale may be presented in a much different form for a larger number of items. However, all forms are the same in the following two respects: The phrases which describe the scale points do not change, and they are always placed so their continuity will be seen.

The method is designed for use with observers who are entirely without experience in food testing. Both the instructions and description of scale points are written with this purpose in mind. However, there is no evidence that the resulting simplicity reduces its effectiveness with other, more sophisticated observers.

The instructions have two functions: One is to tell the observer what he must know, or what the experimenter wants him to know, about the mechanics of the test; the other, and more important function, is to encourage him to report his immediate naive response without any conscious effort to remember or to judge. The simple "like-dislike" description of the scale further encourages this tendency.

Oral instructions can be fully adequate, and they are desirable if the test situation permits individual contact with each observer as in the laboratory. But sometimes the only contact will be through the questionnaire, hence instructions must be carefully written.

No evidence so far available has shown that the geometry or arrangement factors of the scale are very critical, although these have not been intensively investigated. Adjustments are often made for such factors, however, in developing rating scales for other purposes. Laboratory and field forms seem to work equally well. The vertical scale on the laboratory form is to give the idea of a continuum with equidistant points, but there is no proof that this has any definite effect. Apparently most observers consider the laboratory scale only as a series of categories - the same as in the field form.

The lab scale is normally presented with the "like extremely" category at the top or left. But reversal of both the horizontal and vertical scales has been tested and no definite differences have been found. Also, the physical size of the scales has varied from 5 to 7 in. Probably extreme variations in scale size would affect results, although this has not been experimentally determined.

The Taste Tester

Selection of test people is of great importance to interpretation of the results. However, the adequacy of the method does not depend too heavily upon this. The objective is to measure group responses toward foods. Whether or not the group tested represents the consumer group in which we are interested, or whether it represents any group at all, must be determined independently of the test itself.

The number of people required for a given test cannot be arbitrarily stated, but must be determined by the experimenter by the nature and importance of the problem and the degree of precision desired in the results. In the laboratory, the number potentially available usually influences that decision, too.

Variability in ratings from a group of observers tends to be high but also tends to be fairly constant. Thus, it becomes possible to estimate with some accuracy the number of observers necessary to assure statistical significance for a given scale-point separation between two test foods. Because of this high variability, the scale is not suited for use with very small panels—the standard number of observers for tests at the Institute being 40, although this may be increased for important problems. The number of respondents in a field test is usually determined by criteria other than the desired precision of the results.

The tester in the laboratory receives a maximum of three foods at any session. He is instructed to rate each food as he finishes it, and to rinse his mouth with water between samples. Further, he is asked not to change a rating once it has been made. This is done to encourage him to consider the foods independently of one another. Whenever more than one sample is presented it must be assumed that ratings may mutually influence each other.

It has been demonstrated that there is a definite contrast effect when foods being tested lie far apart on a scale. Presenting only one sample at a session would prevent this but is wasteful of laboratory and observer time. Note that later samples cannot affect the earlier ratings except for the occasional observer who fails to obey instructions. Forward-acting effects are equalized by varying the order of presentation of the samples.

The laboratory test situation is designed for optimum sensory discrimination, and the observer's immediate impression is recorded with no opportunity for a memory lag. In field testing, this cannot be done too often. The respondent will usually eat the test food along with other foods as part of his normal meal, and then be asked to comment on it anywhere from a few minutes to several days later.

Thus, when general preference attitudes are surveyed, the delay factor becomes even more important, with the possibility that the respondent may answer from experiences which he could not remember even if he tried. Therefore results of laboratory tests have been found to be more reproducible than those of field tests.

Checking the Data

Two approaches are used in analysis of the data. Each results in a kind of "preference index."

In the first approach, numbers from 1 to 9 are assigned to the scale's nine categories and the data then treated quantitatively. The numbers may begin at either end of the scale, but to have high numbers reflect preference, 9 is usually assigned the "like extremely" end. Score distributions are then dealt with by usual statistical procedures. Calculated are means, standard deviations, standard errors of the means, and the significance of differences between means. And both scores and means may be heated by analysis of variance.

The validity of using certain of the statistical methods with data of this type is questioned by some statisticians. But analytical methods are required and use of the normal procedures is justified on practical grounds until more appropriate techniques become available. The mean rating is the statistic most often used at the Institute, being reported as the major test result.

The second approach provides an index which is probably as useful as the first for the practical purpose of describing a group response to a food. Statistically, it is more respectable, since it deals only with the percentages of responses falling into the various categories. However, unless the number of observers is large, the percentage of responses in some of the categories may be zero, or very close to it. Then it is more convenient and meaningful to combine the categories.

One grouping which has self-evident validity is a combination of the four categories of "dislike." This statistic is always calculated by us and reported along with the mean rating. Also, the percentages falling into single categories, such as "like extremely" or "dislike extremely,"may be useful in special analyses. Results for four foods which were tested at different times in the laboratory are presented in Table I. The entire distribution of responses is shown, along with the statistics which are usually calculated. Food A's rating is unusually low, and that item would be considered nonacceptable under any circumstances; food D has one of the highest ratings ever obtained in the laboratory tests. Of the other two, B would be considered "poor" and C, "good."

Table I - Distribution of Responses on Hedonic Scale, With Resulting Statistical Indices for Various Food Items

		Frequency of Responses			
Scale Point Description	Assigned Value	Food A	Food B	Food C	Food D
Like Extremely	9	0	0	9	11
Like Very Much	8	1	5	12	21
Like Moderately	7	1	11	7	8
Like Slightly	6	8	4	7	0
Neither Like Nor Dislike	5	3	3	2	0
Dislike Slightly	4	4	6	0	0
Disklike Miderately	3	6	6	0	0
Dislike Very Much	2	1	5	2	0
Dislike Extremely	1	0	0	1	0
Total Responses		40	40	40	40
Mean Rating		3.48	5.20	7.08	8.08
Standard eviation		1.99	2.04	1.93	0.68
Percentage "Dislike" Responses		67.5	42.5	7.5	0.0

The broad distributions of responses and the resulting high standard deviations that occur with all of the foods except D are typical. They do not necessarily indicate lack of precision in the method, but reflect the fact that there are normally wide differences among people in their feelings about foods.

Adequate Accuracy

Questions of both theoretical and practical importance should be asked about any new and relatively untried method. These concern its reliability, its precision of discrimination, and its validity for various purposes.

Experimental evidence in regard to the hedonic scale method is far from adequate, but inferences can be drawn from results obtained over its two year period of use in our laboratory This evidence suggests that the mean hedonic rating from as small a group as 40 observers will have satisfactory stability.

However, we may look at the question of stability in two different ways. First, how reproducible is an individual rating, or a mean rating, when a test is repeated under identical conditions?

Three ration items were rated by 35 observers and the test repeated three weeks later without their knowing that the same items were involved. The sets of individual ratings correlated very well and all the mean ratings were reproduced within 0.2 scale points. Under these conditions there seems to be little question about the reproducibility of results.

Table II - Reproducibility of Hedonic Scale
Results, With Groups of 40 Observers Repeatedly
Testing From Foods at Different Times

		Analysis of Mean Ratings				Range/ Standard Error	
Food	Number of Tests	Lowest	Highest	Range	Standard Error[1]	Actual[2]	Chance[3]
Fresh Milk	11	7.02	7.88	.86	.19	4.5	3.4
Lemonade	4	6.60	7.29	.69	.32	2.2	2.6
Canned Bread	8	6.09	7.05	.96	.25	3.8	3.2
Pea Soup	6	7.08	7.65	.57	.22	2.6	2.9

[1] Average of individual stanrard errors of the means.

[2] Figure of Column "Range" divided by figure of Column "Standard Error."

[3] Expected value of ratio if all differences between means were due to chance

Second is the point that for practical work, one wants to know what happens when conditions are not as carefully controlled. And here, even under ordinary conditions, mean ratings have been found to be stable enough to give the impression that a constant property of the food is being measured. Results will generally fall within the narrow range of values permitted by the experimental error.

Table II shows results obtained upon repeated testing of four foods by groups of 40 each. Some of the tests were widely spaced in time, the combinations of foods served were not constant, and the groups were never the same, although they were drawn from the same population. Presumably, the quality of the test foods remained the same, though slight variations may have occurred. Statistical methods were used to find out whether the observed differences between ratings for a food were due to chance. The actual ratio of the range of mean ratings to the average standard error for each food is shown in the column, "Range/Standard Actual" of the table. The expected value of this ratio for any given number of test repetitions, in situations where only chance is operating, may be calculated by statistical methods. These values are shown in the column, "Range/Standard Chance" of the table.

In only one case does the actual ratio appreciably exceed the chance ratio. This suggests that most differences will be accounted for as normal chance variation and that the test is inherently reproducible.

Precision of discrimination with a rating scale is determined by the scale distance between mean ratings, the variability within the distribution of individual ratings, on each item, and the number of observers used in each test. If the scale is measuring hedonic value at all, and if there is a true difference in group response toward the test foods, the difference can be proven simply by increasing the number of observers as necessary.

What Scores Mean

Approximately, 2,000 tests run in the Institute laboratory on over 100 different items have shown a total range in mean ratings from 2.9 to 8.5. Generally, mean ratings below 5.0 represent either poor quality foods or foods that are strange to the observers, while those over 7.5 are obtained for good quality samples of highly popular foods, such as ice cream and candy.

Most foods fall in the range of 5.5 to 7.5, with variability among individual ratings tending to be high. The standard deviations shown in Table I for foods A, B, and C are typical. When variability is of this order and the observer group consists of 40 persons, differences in mean rating of about 0.8 scale units will usually be significant in the sense that they will be reproducible about 95 percent of the time.

Theoretically, then, at least six mutually exclusive levels of hedonic value could be established over the total range.

The hedonic scale rating reflects the attitudes of a group of people toward certain foods and under a given set of conditions. How well the observers and the test conditions represent any practical use situation will depend upon the adequacy of the test plan and the sampling procedures.

Since the hedonic scale method creates no unique problems in this regard and has no special limitations, the factors affecting its validity will not be discussed in detail. However, the ease with which the scale is understood by most people, and its fairly good observed reliability, suggest that in regard to validity for uses involving prediction of consumer preferences it will certainly be as good as other available methods.

Experience to date has shown certain purposes for which the hedonic scale method is valid if the sampling of observers is appropriate and the tests are properly run. These may be summarized as follows:

1. To detect small differences in the direct response to similar foods.

2. To detect gross differences in the direct response to foods, even when time, subjects, and test conditions are allowed to vary.

3. In field questionnaire surveys, to reveal differences in group preference attitudes toward foods.

A fourth purpose may be included, but with some reservation. This is to make general predictions, on an absolute basis, about the acceptance level of any food.

But many people who are not familiar with the problems of preference measurement tend to consider the indices derived from this scale as if they were fixed and unchangeable indicators of acceptance. Available evidence does not bear this out. Even though ratings have a certain stability, they will vary with such factors as the psychological and physiological state of the consumer.

Obviously, they may also vary according to the type of consumer group tested. Thus, it is undesirable to try to establish or use fixed standards of interpretation. For example, mean ratings below 5.0 are usually obtained only for poor quality foods, but it cannot be categorically stated that this number marks the boundary between "acceptable" and "nonacceptable." Exceptions have been found where foods which rate below 5.0 show satisfactory field acceptance.

Another caution is also in order: The hedonic scale method cannot be considered for quality control of flavor in food production. Even though the method has been discussed only in relation to the measurement of preferences, this caution is believed necessary, since the two problems are not always recognized as being essentially different.

Two factors tend to disqualify the method for this type use: 1. Large test variations mean a considerable number of observers are required for precision. 2. The type of responses that are called for are expected to change with a number of conditions which cannot always be controlled. And to correct these would again require a larger number of observers in each test than generally are available for quality control work.

Research on this new method is continuing at thė Institute with these two main objectives: Improvement of the method itself, and determination of the relative importance of the factors that affect hedonic responses toward food.

References

1. **Beebe-Center, J. G.,** *Pleasantness and Unpleasantness,* D. Van Nostrand. New York. 1932.

2. **Guilford, J. P.,** *Psychometric* Method, McGraw-Hill New York, 1936.

3. **Hanson, H., Kline, L. and Lineweaver, H.,** *"Application of Balanced Incomplete Block Design to Scoring of Ten Dried Egg Samples,"* Food Tech., 5, No. 9. 1951.

4. **Grant, B. L.,** *Statistical Quality Control,* McGraw-Hill, New York, 1940.

Development of a Scale for Measuring Soldiers' Food Preferences [a,b]

LYLE V. JONES
University of Chicago, Chicago, Illinois

DAVID B. PERYAM
Quartermaster Food and Container Institute for the Armed Forces
Chicago, Illinois
and

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

[a] This 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.

[b] Presented at the Fourteenth Annual Meeting of the Institute of Food Technologists, Los Angeles, California, June 29, 1954.

Acceptability, always an important consideration in food development and utilization, is particularly so in problems of mass 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.

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 you 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 that exhibits the characteristics of an equal interval scale, the method being based on the assumption that each phrase has a model 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. The 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 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 of the scale and another group placed it two steps above the center. Each instance of a non-normal distribution can be diagnosed as an indication 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 departure from normality.

	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

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
Expecially 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	1.15	.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 intensly	-3.33	1.39
Loath	-3.76	3.54
Dislike extremely	-4.32	1.86
Despise	-6.44	3.62

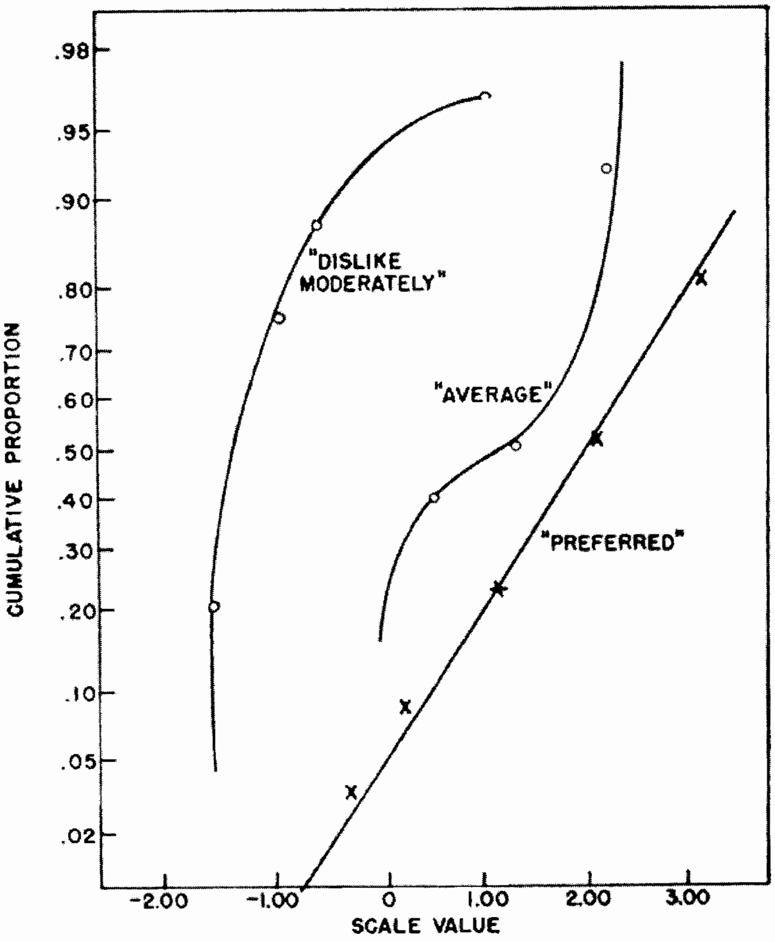


Figure 2. Graphical plots for three descriptive phrases displaying different types of distribution

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 1953. 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.

The following criteria were established to determine the relative adequacy of the scale: (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 (e) 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

Averages of test-retest reliability over 20 food items for nine scale types

Scale Number	Number of Intervals	Characteristics	First Survey	Second Survey
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 midpoint, or neutral catagory, 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

SCALE FOR MEASURING FOOD PREFERENCES

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

Figure 3. Scales investigated in the field surveys

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 8-interval scales are 5-10% higher than those for the 9-interval ones. The 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.

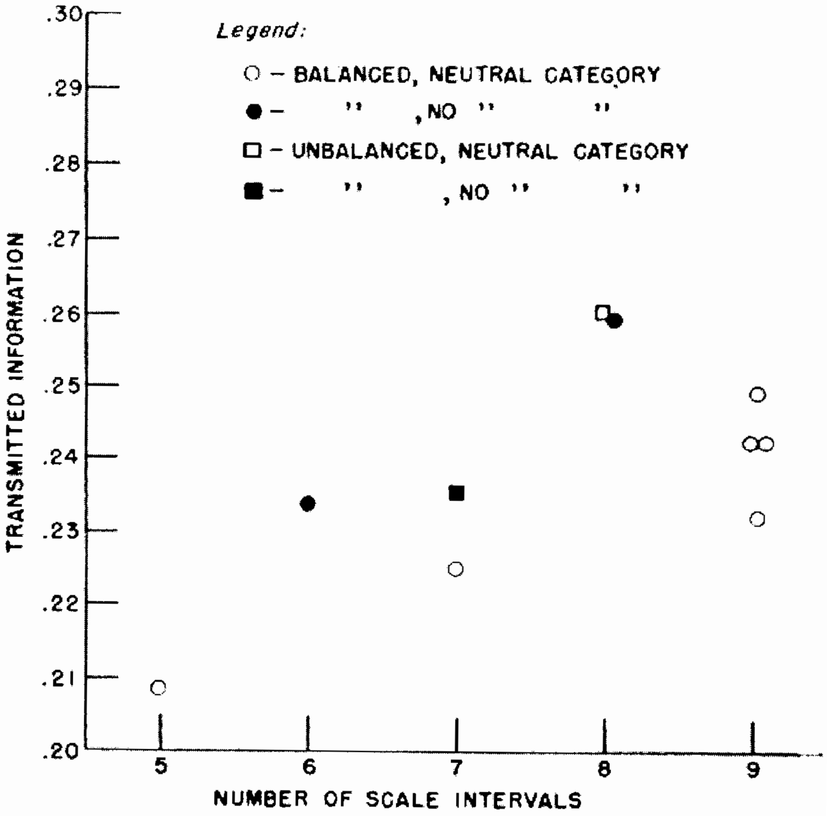


Figure 4. Transmitted information in relation to number of scale intervals.

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 Hume 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 Publg. Co., Inc., New York: 1953.
3. **Garner, W. B., and Hake, H. W**., The amount of information in absolute judgments. *Psychol. Rev.*, 58, 446459 (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, 341 (1953).
6. **Peryam D.R., and Girardot, N.** F. Advanced taste test method. *Food Eng.*, (July, 1952).

Hedonic Scale Method of Measuring Food Preferences

David R. Peryam

and

Francis J. Pilgrim

Quartermaster Food and Container Institute for the Armed Forces
Chicago 9, Illinois

[a]This paper reports research undertaken at the Quartermaster Food and Container Institute for the Armed Forces and has been 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.

[b]Presented at the Seventeenth Annual Meeting of the Institute of Food Technologists, Pittsburgh, Pa., May 16, 1957.

The hedonic scale method has proven to be a very useful tool in food research; however, in no sense is it an "invention" or new discovery. There are only a limited number of basic psychometric methods, and the hedonic scale is no more than a special application of the most generally useful one - the rating scale. It represents a direct approach to the measurement of psychological states.

Background of the hedonic scale method includes the whole history of the development of rating scales, but only a few high-lights can be mentioned. Practical uses of the rating scale date back over 150 years and in those early days included measurement of such things as bath water temperature, wind velocity and other weather phenomena *(3)*. It was therefore on the scene about 70 years before either the rank order or paired comparison methods. The frequency of its use in psychological studies began to increase around the turn of the century. Its development was markedly accelerated after the first World War when psychology began to move out of the classroom and come to grips with practical problems, such as those of education and personnel selection and evaluation. During the last 30 years, and particularly the last 10, the range of psychological applications has increased tremendously and there has been a corresponding increase in the need for psychometric methods. Rating scale methods have been improved through experience and developments in theory. Today they are used extensively in personnel work, consumer research, opinion polling, and in many phases of psychological research.

When were rating scales first used in food evaluation? Scoring and grading systems, which may be considered as forms of the rating scale, were in use by the 1920's *(1)* and may have been applied much earlier. During the past 30 years we have seen many types of rating scales used in food research, technology and quality control.

Essentials of the rating scale method are, first, the definition of a psychological continuum and, second, the establishment of a series of successive categories of response. How good the application is depends upon whether the defined continuum and the categories are meaningful to the test subjects within the context of the problem. The essential features of the hedonic scale are its assumption of a continuum of preference and the direct way it defines the categories of response in terms of *like* and *dislike*.

First work by the Quartermaster Food and Container Institute on this method was done in 1947. A 7-point scale expressed in terms of *like* and *dislike* was used in a survey to determine soldier preferences for menu items. The purpose was to compare this method with the then standard paired comparisons. The rating scale was found to be at least as good and perhaps better. In 1949, when we were seeking a more suitable method to evaluate preference in the laboratory, this idea was re-activated. After a certain amount of preliminary work comparing scale lengths and wording, the present form was selected on the basis jointly of reliability and discrimination. Independent experiments had verified the same form as suitable for questionnaire surveying of soldiers food preferences in the field.

The method was first published in 1952 *(5)*, as an initial attempt that could serve as the basis for further development. It was, perhaps, too immediately successful. We tried it not only for laboratory work but also for attitude surveys *(10)* and various other field situations, with generally satisfactory results. Industry and other Government laboratories also began to use it. As a result, the scale has been changed only for special experiments. The Institute is still using the original form for regular work, even though we have known for some time that better forms could be developed for various purposes *(4)*.

The scale is named for the type of response that it seeks to elicit, i.e., one that derives mainly from feeling or affectivity - in general, the emotional aspects of mental life as opposed to the intellectual. Both science and intuition tell us that the most important dimension in this aspect of human experience is the one indicated in the dichotomy, "pleasant-unpleasant." We can conceive of a continuum lying along this dimension on which, theoretically at least, we could place any emotional experience. "Hedonic" is another, though less common, way of referring to this area. Any scale or test which sought to measure on the same continuum could properly be called "hedonic."

Description of the Method

Basically, what does one need to know to use the method effectively? What are its major advantages and limitations? Figure 1 shows the questionnaire currently used in laboratory preference testing at the Institute. It provides for the rating of up to 4 samples at a session, each one on its separate scale. Presentation is by the single stimulus method, i.e., samples are served individually in succession and each is eaten and rated before the next is served. The test subject is provided with a glass of water and is instructed to "take a drink" during the 40-60 second rest period between samples. Typically, samples are identified by a number or letter code which the subject writes in at the head of the scale.

As for the scale itself simplicity is its essence; little beyond display is needed to put across the idea. The 9 phrases are arranged along a line, or scale, designed to suggest a single continuum which is emphasized by the successive degrees of affect of the verbal description. The instructions are also designed to suggest the continuum and make the subject's task simple. They are printed separately on cards which are posted in the panel booths and are brought to the attention of new subjects. They read about as follows:

"You will be given several servings of food to eat and you are asked to say about each, how much you *like* or *dislike* it. Use the scales to indicate your attitude by checking at the point which best describes your feeling about the food. Keep in mind that you are the judge. You are the only one who can tell what you like. Nobody knows whether these foods should be considered good, bad, or indifferent. An honest expression of your personal feeling will help us decide. Take a drink of water after you finish each sample and then wait for the next.''

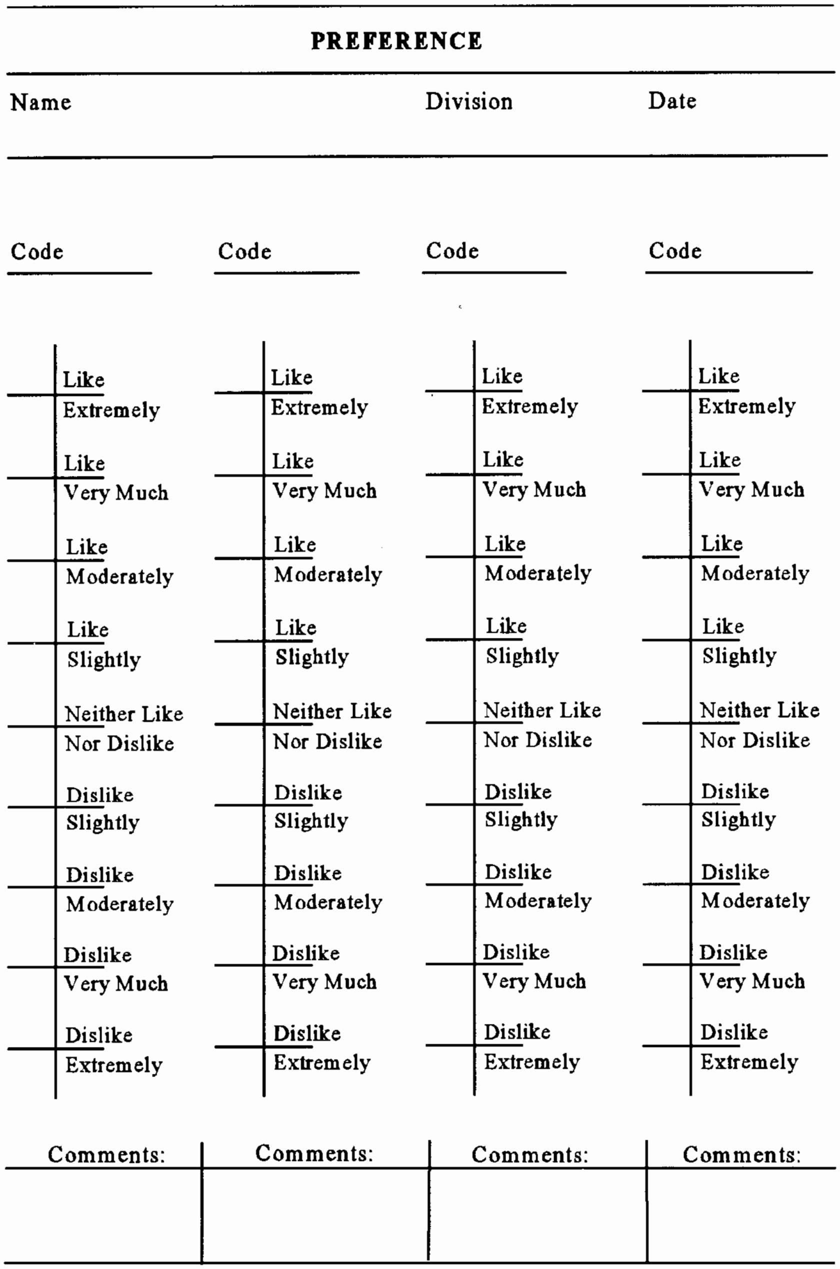
Note that the instructions have two functions: first, to describe the mechanism of the test; second, to encourage freedom of response. The intent is to have the subject answer on the basis of his first impression and to minimize the intellectual approach, i.e., one involving conscious judgement, though, of course these cannot be entirely avoided. The hedonic scale method is predicated on the belief that direct responses, which we may assume are based to a considerable extent on feelings, are more valid for predicting actual behavior toward food than are responses which depend more on reasoning. Both scale and instructions are designed for use with subjects who are entirely without experience in food testing. The "naive" subject responds quite adequately; on the other hand there is no evidence that this simplicity makes the method any less effective with more sophisticated subjects.

The specific way in which the scale, or scales, are presented on the questionnaire, i.e., long or short lines, vertical or horizontal orientation, or whether it begins with *like* or *dislike*, does not appear to be critical. A number of variations have been used with apparently no major effect on the results. They are used not only in the laboratory but also for the preference testing of foods by soldiers in field situations, e.g., when foods are served as a part of a regular meal.

Figure 2 shows the hedonic scale in a different context. This form is employed in questionnaire studies of food preferences and attitudes *(10)*. In the laboratory tests a person is given actual food samples to eat. Hence the sensory components of his experience, based upon the flavor of the specific samples, will be very important. Consequently, he can consider only a few foods at one time. However, we must recognize that his response is determined, in part, by his general attitude toward the food type based on his prior experiences with it. The food preference survey is designed to explore this area. Here the stimuli are not actual food samples, but simply food names. A questionnaire can include up to about 60 items without any evidence that people begin to answer carelessly. Much of the following discussion will apply to both survey and laboratory preference data; however, each has its own special problems of analysis and interpretation. Our primary interests here will be with the sensory test data with occasional references to that from the attitude studies.

Analysis of the Data

The potential of the method can only be realized through appropriate analysis of the data. In most instances this implies that a well-planned design must be set down before the study begins. Scale responses can be treated either as categorical, discrete data or they can be considered as points on a continuum so that the statistics of variables applies. Analysis as discrete data is insensitive compared to the other approach and is used only in special cases. Therefore, only the second approach will be considered here.

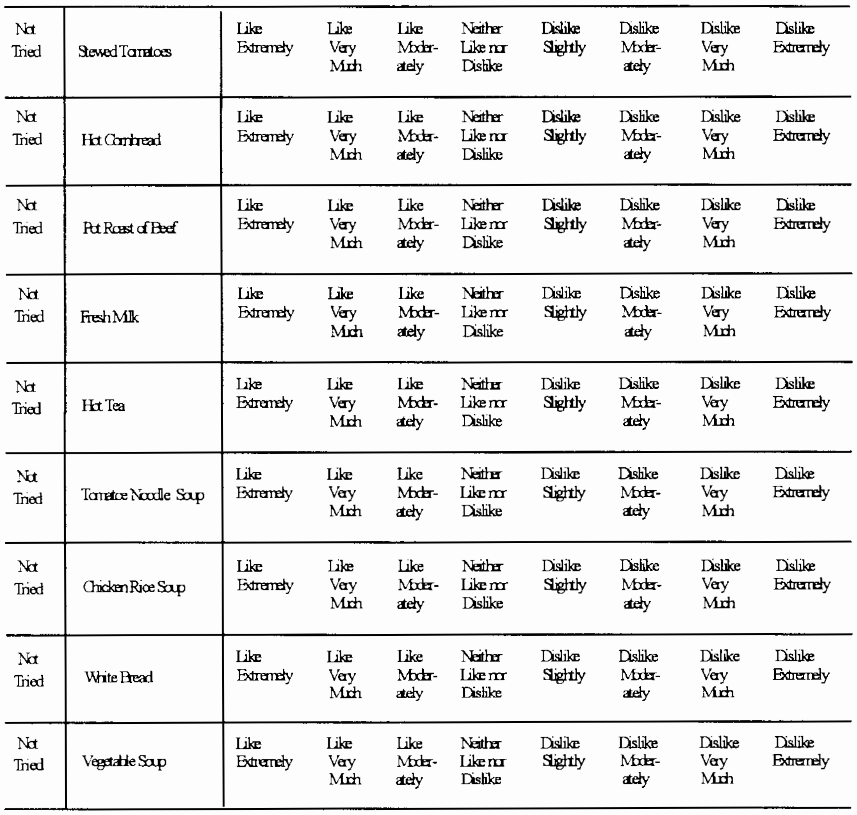


PREFERENCE

Name Division Date

Code ______	Code ______	Code ______	Code ______
___ Like Extremely	___ Like Extremely	___ Like Extremely	___ Like Extremely
___ Like Very Much	___ Like Very Much	___ Like Very Much	___ Like Very Much
___ Like Moderately	___ Like Moderately	___ Like Moderately	___ Like Moderately
___ Like Slightly	___ Like Slightly	___ Like Slightly	___ Like Slightly
___ Neither Like Nor Dislike	___ Neither Like Nor Dislike	___ Neither Like Nor Dislike	___ Neither Like Nor Dislike
___ Dislike Slightly	___ Dislike Slightly	___ Dislike Slightly	___ Dislike Slightly
___ Dislike Moderately	___ Dislike Moderately	___ Dislike Moderately	___ Dislike Moderately
___ Dislike Very Much	___ Dislike Very Much	___ Dislike Very Much	___ Dislike Very Much
___ Dislike Extremely	___ Dislike Extremely	___ Dislike Extremely	___ Dislike Extremely
Comments:	Comments:	Comments:	Comments:

Figure 1. Hedonic scale form used for laboratory preference tests.



Not Tried	Stewed Tomatoes	Like Extremely	Like Very Much	Like Moderately	Neither Like nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely
Not Tried	Hot Cornbread	Like Extremely	Like Very Much	Like Moderately	Neither Like nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely
Not Tried	Pot Roast of Beef	Like Extremely	Like Very Much	Like Moderately	Neither Like nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely
Not Tried	Fresh Milk	Like Extremely	Like Very Much	Like Moderately	Neither Like nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely
Not Tried	Hot Tea	Like Extremely	Like Very Much	Like Moderately	Neither Like nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely
Not Tried	Tomatoe Noodle Soup	Like Extremely	Like Very Much	Like Moderately	Neither Like nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely
Not Tried	Chicken Rice Soup	Like Extremely	Like Very Much	Like Moderately	Neither Like nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely
Not Tried	White Bread	Like Extremely	Like Very Much	Like Moderately	Neither Like nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely
Not Tried	Vegetable Soup	Like Extremely	Like Very Much	Like Moderately	Neither Like nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely

Figure 2. Hedonic scale form used in questionnaire surveys.

One limitation that must be recognized in applying the ordinary statistics of variables is that the scale categories are known to be of unequal width psychologically; therefore, some of the assumptions are not justified. In practice, however, we assign successive integers to the catagories and analyse as though they were equal without running into any major trouble. It has been shown that thr results are quite similar whether one uses integers or the true scale values. This means that all of the modern statistical techniques can be applied.

Most food technology problems are concerned with differences among treatments of the same food. In experiments using the hedonic scale, the treatments may represent either discrete or continuous variables, and there may be more than one variable in a study. The usual statistical tests can be used, such as the *t* test for the difference between two means or the Duncan test *(2)* for differences within a group of means. Variance analysis is appropriate in any case, and is necessary to achieve maximum sensitivity, but, in complex studies, it must be coupled with careful experimental design to give maximum efficiency. Factorial designs are to be recommended in multi-variate problems. For variables involving more than 2 levels, the levels should be distributed equally in terms of some mathematical function. This permits analysis using orthogonal polynomials or curve fitting so that the shape of the curve can be specified and the optimum located. Use of such experimental designs greatly increases the sensitivity of the tests, or alternately, reduces the labor necessary to obtain a definitive answer.

When the question is one of the relationship of preference to a physical or chemical measure - or any other measure such as cost or volume of sales - correlation and regression analysis are applicable.

Reliability

The reliability of a method may be considered in different ways. First is the question of reproducibility of the results when all conditions have remained constant - test-retest reliability. In this sense the hedonic scale is adequate. Good reproducibility, well within statistically estimated limits, is usually found when the same food items have been tested repeatedly under the same conditions. However, another order of reliability is sometimes important to the user of the results, who would like to have a test which gives the same results time after time and regardless of conditions, for example, to have the average rating for a given food vary only within a small range no matter where, how, or by whom it was tested. Considering the many factors which may affect the individual ratings, this is probably too much to expect.

The average rating for a food is affected by 3 sources of variation. Each of these sources may or may not be of concern, depending on the purpose of the experiment. In variance analysis terms, the first kind of variation is the judge-treatment interaction. This is a measure of the disagreement among people as to which treatment is preferred over another and by how much. Such variation is always present, even when the same subjects rate all treatments at the same session, which is the situation for the majority of food technology problems where one is concerned with differences due to processing of raw materials. This variation is minimal; differences in means of 0.4 to 0.5 of a scale point are usually significant when 30 to 40 subjects are used. Also, the relative preferences established in such a test tend to be stable.

The second kind of variation is that due to sampling errors among subjects drawn from the same population. It results both from individual differences in liking for foods and from differences in the way people express themselves. Some are enthusiastic, others restrained; within a given group there may be people who tend to use the low end, others the high end, and still others the middle of the scale. Such variation becomes important when all samples cannot be tested at the same time, and by the same people, as in a storage study.

In this case judge variation must be used as the error term, and the test is usually less sensitive to differences than when the treatments are rated by the same subject. From 0.8 to 1.0 scale point difference between means is usually required for significance at the 5% level. This variation can be called a "session" effect. It is also important when there are more treatments than can be tested by one subject. However, in the latter case, the effect can often be confounded, by proper experimental design, with effects that are of little importance.

The third kind of variation can also be called a session or group effect, but is of a different order and will be added to variation from the other sources. It arises when samples are tested under different environmental conditions, in different test situations, or by different populations. Variation of this kind can be much larger than that from the other sources. It causes differences in the level of rating, e.g., the same people will rate hot soups higher in the winter than in summer and soldiers in the field tend to rate certain canned meats lower than do laboratory panels. It may cause differences in relative preference, e.g., of 2 formulations of chili, people from one population may prefer the spicy and people from another population, the bland. The level of rating seems to be affected more often than the relative order of preference. In one experiment comparing laboratory ratings for 12 ration items with ratings obtained from groups of soldiers in the field, the soldier's average rating was 0.8 of a scale point lower and they showed a 1.0 scale point greater range between items; however, the correlation between the two groups was +.92 *(6)*.

This third kind of variation cannot be considered in assessing test reliability. The fact that the scale will reflect differences in environment, test conditions, or populations does not mean that it is unreliable, but simply shows that it is sensitive. Real changes in preference are being measured. Other types of tests, such as rank order or paired comparisons, do not pick up these changes in preference level. They provide for comparison only among samples that are simultaneously present; hence they give only relative preference.

Essentials of the Method

Certain essentials of the hedonic scale method, such as the single stimulus method of presentation, the establishment of a well-defined continuum, and the selection of phrases to make the intervals as clearly successive as possible, are standard features of the rating scale methods. The real contribution of the hedonic scale method lies in application of these principles to the problem of measuring food preferences. Briefly summarized, the essential features of successful application are:

A) definition of the continuum as one of affectivity, rather than judgement. B) structuring the scale with like and dislike terms, thaat are easily understood and meaningful, and C) the policy of not "tampering" with the subjects but encouraging free, uninhibited expression. Other characteristics of the scale are not considered essential to the method. The number of scale categories may be changed without basically changing its function. Research at the University of Chicago *(4)* has shown improved discrimination up to 9 categories, with some indication that a larger number might be even more effective. It is not critical to have the category descriptions exactly as in the present form. They must, of course, clearly indicate the affective continuum and encourage its use, but this might be done as well with other words. Further, some phrases and series of phrases are more effective than others in getting across the idea of the successive nature of the scale intervals. Ambiguous phrases and words must be avoided. The present scale appears adequate in this regard with the exception of the phrase dislike moderately, which different people seem to understand differently. Apparently the scale would discriminate better if a less ambiguous phrase could be substituted for this one. Another feature shown to be non-essential is the inclusion of the neither like nor dislike category. Its removal would do no harm and might even improve discrimination. It was also indicated that the scale need not be balanced, i.e., it does not have to have an equal number of like and dislike categories.

Interpretation

There are a number of problems involved in the interpretation and use of hedonic scale data. When a food technologist runs a test he does it to find out something about the foods - their characteristics, their probable acceptance by consumers, etc. With his attention thus oriented, he may lose sight of a very pertinent fact. The hedonic scale is designed to measure human behavior potential, not characteristics of food. Foods are evaluated indirectly by making inferences from the behavioral measures. The foods themselves constitute only one of the factors that are contributing to the final test results. Potentially a person may respond to his whole environment, both external and internal. Any aspect of it may control his behavior to some degree at any given moment, and we cannot always predict what part is going to exert the strongest control. In a food preference test, we are sure that the food samples, which may be held constant for all subjects and for all successive tests, are a very important aspect of the environment. But peoples' responses will also be affected by personality, long and short range attitudes, physical well-being, and other physiological and psychological variables *(7)*. Thus, we should expect a test to be sensitive to changes in many factors to a greater or lesser degree.

Special Effects

Several effects which appear in hedonic scale tests have been frequently confirmed and need constantly to be considered in test procedure. They are all concerned with inter-effects among samples when more than one is presented in the same test session.

The first is the so-called "contrast" effect, where an average quality sample will tend to rate low when preceded by a good quality sample of the same type and to rate high when preceded by a poor quality sample. This is seen as a logical result when we consider that the first sample, because of its recency and vividness, becomes an important part of the total frame of reference in which the subject responds to the second. This effect is not observed, or is noted less often, when 2 foods of different kinds are presented. Just to assure that we do not try to make ''contrast'' into an eternal principle we should note a contradictory effect which appears occasionally. We might call it "contamination'' effect. Here the rating for an average quality sample will tend to move in the same direction as those with which it appears. This has been noted in storage studies where the rating of a constant quality control has progressed steadily downward from one storage period to the next as the stored samples deteriorated. Psychology might consider this as a special case of stimulus generalization; the presence of a number of poor quality samples in the series develops a low-preference attitude which affects the better quality sample.

A third effect can be called "position" effect. The first sample presented occupies the best position, preference-wise, and the later samples tend to be rated lower. This effect appears so frequently that it must always be anticipated. However, it varies considerably among foods both as to degree and where in the series it first appears. With some foods, e.g. coffee, the drop in preference occurs after the first sample and is rather large. With others it may he only slight or may not appear at all up to the third or fourth sample. Sometimes, e.g., in testing fresh milk or orange juice, it may be entirely absent. This effect can be balanced out by using all possible orders of presentation of samples an equal number of times.

Applications

The hedonic scale method is flexible. Its use seems feasible in a broad range of situations and with any problem where we might want to evaluate on the criterion of human preference. It has been used most often in laboratory studies where the objective is to obtain information about probable acceptance as a guide to further development work.

Since, as with other methods, relative preference is assessed most accurately and reliably, it has its greatest value when treatments can be compared in the same session. It is used constantly in the Quartermaster Food and Container Institute laboratory to determine the effects of variables such as processing, formulation, raw materials, packaging, and storage conditions. The method is also employed in problems where one must rely on the constancy of the absolute level of rating, such as evaluating the effects of storage over a period of time, although here the error is larger, and conclusions are not so reliable. The absolute level of rating also serves as a preliminary estimate of field acceptance.

Evaluation of food quality, as by a trained panel of product specialists, is often considered as different from consumer preference evaluation. It is assumed that trained panels use derived criteria which are somehow different and perhaps more stable than those employed by the ordinary consumer. However, our experience has been that a trained panel is just another sampling of consumers expressing their preferences. They will give the same results, within experimental error, using the hedonic scale as with any of the special methods which depend upon judgments of quality.

Suitability of the hedonic scale method in a wide range of situations and with many different populations of test subjects derives from its simplicity. The subjective continuum on which it measures is a universally meaningful one and the language it employs is easily understood.

Validity

Test construction doctrine requires consideration of test validity, something which is often neglected in food technology where we are usually satisfied merely to achieve reliability. However, let us assess the validity of the method under discussion by trying to answer the question, "How well does it measure what it is supposed to measure?" First we must decide what it is supposed to measure. A test may be valid, of course, for one purpose and entirely invalid for another.

It is often assumed that the hedonic scale measures acceptance. For example, the statement, ''Food x scored over 5.0 and is therefore acceptable," not only makes this assumption but further assumes that there is a direct, known relationship between scale values and acceptance. Such over-simplification serves no useful purpose, and does harm in diverting one from the attempt at meaningful interpretation. But what does the scale measure? The best answer is that it measures something which we call "preference.'' which is a short-hand method of referring to the hedonic continuum. Preference is measured for the purpose of predicting acceptance, which is tantamount to establishing acceptance as the major validating criterion. Now, how does one define and quantify acceptance! It is not on an all-or-none-phenomenon - you cannot say that people either do or do not accept a food. There are degrees of acceptance. Some objective indices of acceptance that have been used, or suggested, are amount of food consumed, frequency of choice of a food among competitive foods, and, in the market scene, frequency of purchase. One definition of acceptance combines the objective and subjective and affirms that we should consider it as ''consumption with pleasure'' *(7)*. Most of us would intuitively agree that the ingestion of food serves a broader purpose than simply nourishing the body.

Let us, then, accept both the measurement of preference and the prediction of acceptance as the objectives of hedonic scale measurement. How well does it do either? To get an objective validating criterion for measurements on the affective continuum is difficult but not impossible. Physiological and behavioral indicators might be found; however, the experience has been that such measures are less precise and reliable than are people's verbal statements. Further, their theoretical justification is no better - if as good. Essentially, then, we must rely on the face validity of the hedonic scale as a good measure of preference. We accept it because it obviously is measuring what it is supposed to measure, and because there is no better measure against which to check it.

This same argument applies to the "pleasure" part of the "consumption with pleasure" definition of acceptance. However, we may still ask about the prediction of consumption or choice. We know that many things, in addition to preference as measured in any test situation, will affect choice and consumption. Some of the major influences are ideas about health and nutrition, cost, the need for calories as affected by climate, activity, and other factors, and the availability of competitive foods. Data are available from studies conducted by the Quartermaster Food and Container Institute that show that preference is very important in spite of the many other factors that must be considered. In an extensive experiment involving the feeding of 100 soldiers over a one-month period in a special mess-hall situation where they had considerable freedom of choice in regard to the foods they could select, it was shown that preference ratings obtained by the questionnaire method correlated about +.74 with both the frequency of item selection and the amount consumed *(9)*. Other investigations where consumption or frequency of rejection were correlated with preference ratings have given comparable results. Depending on conditions and type of food the correlations have varied between +.30 and +.87. We may conclude that, generally, over half of the variation in these objective indices is explainable on the basis of preference.

Conclusions

The most serious limitations of the hedonic scale method are those that apply equally to any measurement of preference under restricted conditions, such as inadequacy of the sampling of test subjects, the possibility of bias, and the fact that tests are run under only a limited range of conditions that may or may not be appropriate. Certain other limitations are inherent in the rating scale method, such as its susceptibility to the various "effects," rather high variability in the data, inequality of scale intervals, and the lack of a zero point. In deference to those whose expectations are highly optimistic, it is noted that values on the scale cannot validly be interpreted in terms of objective food acceptance behavior except within a wide range of error. However, in general these are limitations only when physical measurement is used as the standard of comparison. Whether or not there are limitations in comparison with other methods of measuring preference is still being debated. The suggestion has been made that it does not discriminate as finely between samples as paired comparisons. This has been disproven for foods in the average preference range *(8)*; however, the possibility that paired comparison does discriminate better with well-liked foods is still open. Major advantages of the method are: (a) its simplicity, which makes it suitable for use with a wide range of populations, (b) subjects can respond meaningfully without previous experience, (c) the data can be handled by the statistics of variables - an advantage inherent in rating scale data, and (d) in contrast to other methods, within broad limits the results are meaningful for indicating general levels of preference.

LITERATURE CITED

1. **Boggs, M .M. and Hanson H. L.**, Analysis of food, by sensory difference tests. In *Advances in Food Research*, Vol.II. Academic Press, Inc. New York. 1949.
2. **Duncan, D. B**. Multiple range and multiple *F* tests. *Biometrics*, 2, 1-42, 1955.
3. **Guilford, J. P.** *Psychometric Methods*. McGraw Hill, New York, 2nd ed., 1954
4. **Jones, L.V., Peryam, D. R. and Thurstone L. L.** Development of a scale for measuring soldiers' food preferences, *Food Research*, 20. 512-520, 1955.
5. **Peryam, D. R. and Girardot N. F**. Advanced taste test method. *Food Eng*. 24(7), 58-61, 194 (1952).
6. **Peryam D. R. and Haynes J. G**. Prediction of soldiers' food preferences by Laboratory methods. J. *Applied Psychol*, 41, 2.6, 1957.
7. **Pilgrim, F. J.** The components of food acceptance and their measurement. *Am. J. Clin. Nutr.*, 5, 171-175, 1957.
8. **Pilgrim, F. J. and Wood, K. R.** Comparative sensitivity of measuring consumer preference. *Food Technol*. 9 , 385-387,1955.
9. **Schutz, H. G.** Preference ratings as predictors of food consumption. *Am. Psychologist*, 12, 380. 1957.
10. **Wood, K. R. and Peryam, D. R.** Preliminary analysis of five Army food preference surveys. *Food Technol*. 7, 248-249, 1953

Hedonic Differences as a Function of Number of Samples Evaluated

D. R. PERYAM,

D. B. PERYAM
and

B. J. KROLL

Peryam & Kroll Research Corporation
Chicago, Illinois

SUMMARY— This study investigated the limit of the number of samples that can be reliably evaluated for preference in a single laboratory test session. Multiple samples were presented to 200 subjects randomly assigned to five conditions. Two critical samples were included under each condition and the number of other samples varied from zero to four. Two replications of each set were evaluated by each subject in a single session without his knowledge that samples were repeated. The total number of samples for the five experiments were 4, 6, 8, 10 and 12. The experiment was run separately with milk, soup and gravy base and maple syrup. There was no evidence that serving up to 12 samples in a single session adversely affected preference discrimination

Introduction

In the typical laboratory situation, the time a subject actually spends in the taste-test booth is only a part of the total time he devotes to testing. Indirect time costs - breaking his job routine, walking to and from the laboratory, waiting to test or partaking of refreshments - taken together are probably greater than the direct time costs. Thus, it may be assumed that the more samples an individual can evaluate in one session, the more work the laboratory can produce with a given expenditure of subject time. Also, laboratory personnel can be more efficient because of economies in preparation of samples and administration. In central location testing, where paid subjects are recruited from outside the organization, costs can be reduced substantially by getting more information from each person (Girardot, et al., 1968).

Perhaps more important is that often, when each person has tested all samples, more straightforward experimental designs can be used instead of complicated ones such as balanced incomplete blocks. More precise estimates of inter-judge variability and judge-sample interaction can be used in the analyses of variance and the tedium and assumptions inherent in block adjustments can be avoided. Intuitively, there is a limit on the number of samples that can be presented. Most people can be induced to cooperate, but does acquiescence bring with it a loss in performance? Deterioration would not necessarily be due to sensory adaptation and loss of acuity, but might be a function of motivation as reflected in reduced attention, carelessness and general confusion.

Bradley, et al., (1954) reviewed much of the earlier literature on this topic. From two to eight samples per session have been advocated by various researchers, but usually there was no evidence that the recommended number was actually the optimum. These experiments as a group were inconsistent in the ranges of session length investigated. Some started with the assumption that the permissible maximum is two samples, but dared to go as high as three or four. Others went far beyond this limit.

Many experiments have dealt with sensory discrimination rather than with preference. Laue, et al. (1954) reported a loss of discrimination in the second of two successive triangle tests when maple syrup was the test material: however, this was not true for coffee. Mitchell (1956) found no loss of discrimination in duo-trio testing of beer even when the session was extended to four sets. Brandt, et al. (1956) went even further. Using various alcoholic beverages as test media, he tried as many as six duo-trios at one session, and reported that there was no loss of sensitivity. Pfaffman. et al. (1954) had subjects test repeatedly during 40 min sessions, employing the duo-trio and triangle tests with various food types, and found no significant loss of discrimination even after as many as 75 samples.

Preference testing may present a different kind of problem. Even though a subject's acuity might not be affected by continued exposure to food stimuli, his feelings may be altered. Bradley, et al. (1954) found no effect on levels of rating for various foods when they were served early vs late in an eight-sample series. Sather. et al. (1960) had untrained subjects rate a series of 20 samples in four sets of five samples each. No loss of discrimination was found. Girardot, et al. (1968) reported that consumers' preference discrimination was better with the second pair than with the first pair of coffees.

In most preference testing the absolute levels of rating are of secondary importance. The main concern is with the direction and amount of the differences among competing samples. The present study took this approach in investigating the effect of extended testing. The basic question was: do differences in preference among food samples of the same subclass remain constant as the number of samples is increased?

Experimental

There were three replications of the basic experimental design which differed only in the products tested - milk, syrup and a broth made from soup and gravy base. The testing was done in the Food Acceptance Laboratory, U.S. Army Natick Laboratories. according to their normal practices. Samples were evaluated by ratings on a 9-point hedonic scale.

Subjects

Test subjects were selected randomly from the Natick Laboratory taste-test pool of about 700 persons. For each replication, five groups of 40, one for each experimental condition, were selected without replacement. Thus, no subject tasted a given food type more than once. Selection of the samples of subjects for each of the three replications was independent. Thus, it is possible that some people participated two or even three times in al!.

Conditions

There were five conditions which varied only in the number of samples presented. Each condition was run in a separate session. In each session a subject evaluated all samples twice. Each time, subjects were told beforehand the number of samples but not of the duplications. The samples included in the various conditions were: first—A and B: second—A, B and C; third—A, B, C and D; fourth—A, B, C, D and E; and fifth— A. B, C, D, E and F. After rating all samples once, a subject immediately rated them again. The serving orders in each half of a condition were balanced insofar as possible, and the orders in the two halves were independent.

Selection of Samples

Since the analysis focused on the difference between A and B, it was important that these two samples be neither too far apart nor too close together in preference. If the difference were very large, then the maximum possible difference in rating might be obtained even in the least suitable condition. If the difference were too small, then even the most sensitive test method might not be able to demonstrate an effect and the negative results would be meaningless. Hence, the strategy was adopted of pilot testing the series of six samples to be used in the final test and designating the one with the second highest rating as Sample A and the one with the second lowest rating as Sample B.

This plan was carried out for soup and gravy base and milk, but revision was required for syrup. One of the original six samples was sorghum, which had a very low pretest rating (3.77). To have included this item might have induced contrast effects that would reduce discrimination among the other samples. Since it would have been tested in the 6-sample condition, the main effects of numbers might have been confounded with contrast effects in some complex manner. Hence, it was eliminated and the third-ranked sample was duplicated as both D and F. Another change was to designate the best syrup sample as A and the next best as B. since this contrast appeared to suit the purposes of the study. Table 1 describes the samples used in each replication, and gives certain details about the testing procedures.

Results

The results are reported separately by replication. Tables 2 - 4 show, for the three products, respectively, the mean rating of each sample in each condition by experimental half, then the totals. Table 5 demonstrates the effect of the experimental conditions on the ratings of the critical samples, A and B, for all three products. It gives the averages for the first and second halves and the differences between these averages. It also shows the differences between samples (A-B) for each experimental half, and, finally, the differences of the differences (second half minus first half). Note that a positive value in the last column indicates that the second half was more discriminating, while a negative value shows better discrimination in the first half.

Table 6 presents the analyses of variance for the critical samples for all products. Table 7 summarizes the most important information obtained from the analysis of variance made for each condition separately. Only the main effects of sample and experimental half, and the sample-half interaction, are given.

Table 1 - Test samples and serving conditions

Product	Sample description	Pretest rank order	Serving conditions
Soup and gravy base	A (All special formulations)	2	Temperature: 150-155° F
	B	5	
	C	3	Sample: 2 ounces
	D	4	
	E	1	Interval: 30 seconds
	F	6	
Milk	A 92% fresh, 8% recon.[1]	2	Temperature: room
	B 68% fresh, 32% recon.	5	
	C 84% fresh, 16% recon.	3	Sample: 1 oz
	D 76% fresh, 24% recon.	4	
	E 100% fresh, 0% recon.	1	Interval: 30 sec.
	F 60% fresh, 40% recon.	6	
Syrup	A Commercial	1	Temperature: 115-120° F
	B Commerical	2	
	C Pure maple	4	Sample: 1/2 oz
	D Govt. standard[2]	3	
	E Commercial	5	Interval: 60 sec.
	F Govt. standard[2]	3	

[1] Reconstituted - 33% condensed milk and 67% water.

[2] Samples D and F were the same.

Table 2 - Soup and Gravy Base. Mean ratings of samples in each experimental condition (N = 36 in 6-sample, 40 in all others)

	Experimental condition				
	2 Sample	2 Sample	2 Sample	2 Sample	2 Sample
Sample A					
1st half	7.13	6.93	6.68	6.83	6.58
2nd half	6.70	6.98	6.45	6.33	5.81
Total	6.91	6.95	6.56	6.58	6.19
Sample B					
1st half	7.13	6.38	6.28	6.13	5.97
2nd half	6.70	5.58	5.88	5.48	5.14
Total	6.91	5.98	6.08	5.80	5.40
Sample C					
1st half	-	6.58	6.90	6.93	6.28
2nd half	-	6.05	6.23	6.63	6.25
Total	-	6.45	6.56	6.78	6.26
Sample D					
1st half	-	-	6.80	7.33	6.86
2nd half	-	-	6.80	6.68	7.53
Total	-	-	6.80	7.00	6.69
Sample E					
1st half	-	-	-	6.88	6.83
2nd half	-	-	-	6.63	6.06
Total	-	-	-	6.75	6.44
Sample F					
1st half	-	-	-	-	4.56
2nd half	-	-	-	-	3.81
Total	-	-	-	-	4.18
Total					
1st half	6.99	6.72	6.66	6.82	6.23
2nd half	6.59	6.20	6.34	6.35	5.56
Total	6.79	6.46	6.50	6.58	5.86

Soup and Gravy Base

Ratings were obtained from only 36 subjects for the 6-sample condition. Thus, for purposes of the analyses presented in Tables 5 and 6 the ratings given by four randomly selected subjects were eliminated from each of the other conditions to equalize the *N*.

Across the entire experiment Samples A and B differed by 0.61 scale points (Table 5), a difference significant at the 0.1% level (Table 6). Samples were rated consistently higher in the first half than in the second half of the sessions by an average of 0.44 scale points. The main effect of condition was significant at the 1% level. Most of this effect was probably attributable to the 6-sample condition, where the samples were rated particularly low, since the effect was not significant in a separate analysis that excluded the 6-sample part. Note that the average ratings tended to decrease as the number of samples increased (Table 2).

No interaction involving sample, half or experimental condition was statistically significant (Table 6). We would conclude that, although ratings tended to be lower in the second half than in the first, there was no evidence that the differences between the samples were affected either by the number of samples or by whether they were presented in the first half or in the second half of the session.

The summary of the separate analyses by condition (Table 7) shows that Samples A and B always differed significantly (0.1% level) with the exception of the 2-sample condition. Also, the main effect of experimental half was always significant. The interaction of sample and half was significant only for the 3-sample condition (5% level). For this condition, the difference between the samples was much greater in the second than in the first half (1.40 vs. 0.55); however, this increased differentiation failed to appear in the other four conditions. Thus, again, there was no evidence that either the total number of samples or position in the serving order had any consistent effect on the difference between the critical samples.

Milk

The difference of 1.08 scale points between Samples A and B (Table 5) was larger than was anticipated on the basis of the pilot test results. Again, ratings in the first half were significantly (5% level) higher than in the second half (Table 6). The absence of a significant interaction between sample and half implies that the difference between Samples A and B in the first half (0.97) was not significantly smaller than the difference in the second half (1.21).

The main effect of condition was not significant, so there is no evidence that the total number of samples affected the level of rating. Since none of the interactions was significant, we cannot conclude that the total number of samples or position in the serving order affected the difference between Samples A and 13. The figures in Table 5 would seem to indicate otherwise.

The difference between halves was 0.60 for the 2-sample condition and dropped successively to -0.08 for the 6-sample condition, but the sample-half-condition interaction was not significant: hence, this apparent trend has no statistical support.

The results of the separate analyses for each condition (Table 7) maintain a consistent pattern. The main effect of sample was highly significant (0.1 % level) in each case, but none of the sample-half interactions was significant. The only other significant source of variation in any analysis was the main effect of half in the 6-sample condition (1 % level).

Table 3 - Milk. Mean ratings of samples in each experimental condition (N = 40 each condition

	Experimental condition				
	2 Sample	2 Sample	2 Sample	2 Sample	2 Sample
Sample A					
1st half	6.78	6.65	6.75	6.90	6.58
2nd half	6.73	6.75	6.75	6.68	6.38
Total	6.75	6.70	6.75	6.79	6.48
Sample B					
1st half	5.95	5.67	6.08	5.75	5.35
2nd half	5.30	5.28	6.00	5.45	5.23
Total	5.63	5.48	6.04	5.60	5.29
Sample C					
1st half	-	6.23	6.63	6.30	6.08
2nd half	-	6.45	6.33	6.15	5.82
Total	-	6.34	6.48	6.23	5.95
Sample D					
1st half	-	-	6.45	5.90	5.85
2nd half	-	-	6.13	5.60	5.55
Total	-	-	6.29	5.75	5.70
Sample E					
1st half	-	-	-	6.90	6.75
2nd half	-	-	-	6.58	6.35
Total	-	-	-	6.74	6.55
Sample F					
1st half	-	-	-	-	5.45
2nd half	-	-	-	-	4.63
Total	-	-	-	-	5.04
Total					
1st half	6.36	6.18	6.48	6.35	6.01
2nd half	6.01	6.16	6.30	6.09	5.66
Total	6.19	6.17	6.39	6.22	5.83

Syrup

Samples A and B were significantly different (0.1% level) in preference (Table 6); however, one finding was totally unexpected. In the pilot test, Sample A had rated 0.57 scale points higher than Sample B. In the main experiment Sample B was rated an average of 0.34 scale points higher than Sample A and there were only two instances out of ten where Sample A was higher in any half in any condition (Table 5). There was some evidence that the actual formulation of Sample B, a commercial product, had been changed between the pilot test and the main experiment.

As in the other replications, the effect of first vs. second half was significant (5% level), with the second half rated lower. Here the effect of experimental condition was also significant (5% level). The average of A and B dropped from 6.90 in the 2-sample condition to 6.00 in the 6-sample condition (Table 4), but the decrease was not monotonic. The only other significant effect (5% level) was the interaction between half and condition. The right-hand portion on Table 5 illustrates this, but there is no clear reason why it occured. In the first half, the difference between the samples was highest for the 3-sample and 5-sample conditions; in the second half, it was highest for the 2-sample and 5-sample conditions and almost zero for the 4-sample and 6-sample conditions. There was no linear trend for the difference to vary according to the number of samples tested, whether one considers each half separately or both combined.

The absence of an interaction between samples and conditions, or between samples and half, is consistent with the results from the other two replications. Thus level of ratings might be affected by whether samples appeared in the first or second half; but, on an overall basis, the conclusion that Sample B is preferred to Sample A does not depend upon the half in which they were tested or upon the number of samples with which they were evaluated or upon the interaction of the two variables.

The separate analyses by experimental condition (Table 7) again clearly show differences between the critical samples and between halves. In each case Sample B rated higher and level of rating was higher in the first half of the sessions. In the 4-sample and 5-sample conditions, interactions between sample and half were significant. In the 4-sample condition, Sample C dropped by 1.32 scale points from the first to the second half, and Sample D dropped by 1.20 scale points (Table 4). However, the mean rating of Sample A was only 0.27 scale points lower in the second half, and the mean rating of Sample B was only 0.08 scale points lower. Similarly, in the 5-sample condition, Samples C and D dropped 1.27 and 1.37 scale points, respectively, in the second half: but Samples A and B dropped only 0.05 and 0.10 scale points. Thus, the nature of the interaction between sample and half was nearly the same in each of the two conditions.

The four samples which appeared in the 4-sample and 5-sample conditions also were present in the 6-sample condition, yet the sample-half interaction was not significant. Why not? Note (Table 4) that mean ratings for Sample C and D were substantially lower in the first half of the 6-sample condition than in the 4-sample condition and to a lesser extent than in the 5-sample condition. The two samples did drop in the second half of the 6-sample condition by 0.67 and 0.65 scale points, but for some reason Sample B dropped by an unusually large amount, 0.45 scale points, relative to the preceding two conditions. Thus, the absence of a sample-by-half interaction in the 6-sample condition may have been due to abnormally low first-half ratings for Samples C and D or to the abnormally large decrease for Sample B. Either or both of these effects would work in the direction of supporting the null hypothesis for the sample-by-half interaction.

Table 4 - Syrup. Mean ratings of samples in each experimental condition (N = 40, each condition)

	Experimental condition				
	2 Sample	2 Sample	2 Sample	2 Sample	2 Sample
Sample A					
1st half	7.03	6.75	6.60	6.28	5.90
2nd half	6.25	6.38	6.33	6.23	5.90
Total	6.64	6.56	6.46	6.25	5.90
Sample B					
1st half	7.38	7.38	6.43	6.95	6.33
2nd half	6.93	6.60	6.35	6.85	5.88
Total	7.15	6.99	6.39	6.90	6.10
Sample C					
1st half	-	5.45	5.30	4.75	3.55
2nd half	-	4.70	3.98	3.48	2.88
Total	-	5.08	4.64	4.11	3.21
Sample D					
1st half	-	-	5.93	5.70	5.55
2nd half	-	-	4.73	4.33	4.90
Total	-	-	5.33	5.01	5.23
Sample E					
1st half	-	-	-	4.88	5.23
2nd half	-	-	-	4.35	4.80
Total	-	-	-	4.61	5.01
Sample F					
1st half	-	-	-	-	5.47
2nd half	-	-	-	-	4.80
Total	-	-	-	-	5.14
Total					
1st half	7.20	6.53	6.06	5.71	5.34
2nd half	6.59	5.89	5.35	5.05	4.86
Total	6.90	6.21	5.70	5.38	5.10

Even in the two conditions where the sample-by-half interaction was significant, the rank order of preference within half remained the same: Sample B always had the highest mean rating. Sample A the second highest. Sample D next and Sample C the lowest. The range of ratings was somewhat higher in the second half than in the first half: 2.37 vs 1.30 for the

4-sample condition, and 3.37 vs 2.20 for the 5-sample condition. There is no evidence that the differences among samples are attenuated either in the second half or in the experimental conditions involving a larger number of samples.

Table 5 - Effect of experimental condition on ratings of the critical samples A and B

	Average (A + B)/2)			Average (A - B)		
Conditions	1st Half	2nd Half	Difference 1st-2nd	1st Half	2nd Half	Difference 1st-2nd
Soup & gravy base (N = 36)						
2-Sample	7.02	6.56	0.46	0.19	0.23	0.04
3-Sample	6.72	6.35	0.38	0.55	1.40	0.85
4-Sample	6.44	6.24	0.20	0.50	0.47	- 0.03
5-Sample	6.42	5.90	0.52	0.67	0.69	0.02
6-Sample	6.12	5.48	0.64	0.91	0.67	- 0.24
Total	6.54	6.10	0.44	0.55	0.66	0.11
Average of halves	6.32			0.61		
Milk (N =40)						
2-Sample	6.36	6.01	0.35	0.83	1.43	0.60
3-Sample	6.16	6.02	0.14	0.98	1.47	0.49
4-Sample	6.42	6.38	0.04	0.67	0.75	0.08
5-Sample	6.32	6.06	0.26	1.15	1.23	0.08
6-Sample	5.96	5.80	0.16	1.23	1.15	- 0.08
Total	6.24	6.05	0.19	0.97	1.21	0.24
Average of halves	6.14			1.08		
Syrup (N = 40)						
2-Sample	7.20	6.59	0.61	0.35	0.68	0.33
3-Sample	7.06	6.49	0.57	0.63	0.22	- 0.41
4-Sample	6.52	6.34	0.18	-0.17	0.02	0.19
5-Sample	6.62	6.54	0.08	0.67	0.62	- 0.05
6-Sample	6.12	5.89	0.23	0.43	- 0.02	- 0.45
Total	6.70	6.37	0.33	0.38	0.30	- 0.08
Average of halves	6.54			0.34		

Discussion

All three replications (food items) yielded similar interpretations. First, the ratings in the second half were significantly lower than in the first half. This point is inconsequential when considering one major purpose of taste-tests, which is to determine differences among samples rather than to establish levels of rating. However, some users of the data might be pleased or dismayed at the lower ratings in the second half and might have to shift their frames of reference when using such ratings as absolutes.

Second, within each replication the samples were intended to differ sufficiently to allow the effects of other variables to come into play. An unanswered question is whether the differences were, in fact, too great so that they obscured the effects of these other variables. This may have occurred for milk, where the overall mean difference between Sample A and Sample B was 1.08 scale points (Table 5); but the differences of 0.61 for soup and gravy base and of 0.34 for syrup approach the specifications for this experiment. Because the results correspond so well among the three

Table 6 - Analysis of variance for critical samples (A and B) across experimental conditions

Source of variation	df	ms	F
Soup & gravy base (*N* = 36)			
A-Sample	1	66.61	35.44[1]
B-Half	1	35.11	30.27[1]
C-Condition	4	20.16	3.65[2]
A X B	1	0.51	[3]
A X C	4	2.57	1.37
B X C	4	0.98	[3]
A X B X C	4	1.28	1.58
D-Subject (within C)	175	5.53	
A X D	175	1.88	
B X D	175	1.16	
A X B X D	175	0.81	
Milk (*N* = 40)			
A-Sample	1	236.53	123.19[1]
B-Half	1	7.41	6.18[4]
C-Condition	4	5.56	[3]
A X B	1	2.77	2.72
A X C	4	1.81	[3]
B X C	4	0.56	[3]
A X B X C	4	0.88	[3]
D-Subject (within C)	195	6.82	-
A X D	195	1.92	-
B X D	195	1.20	-
A X B X D	195	1.02	-
Syrup (*N* = 40)			
A-Sample	1	23.46	13.72[1]
B-Half	1	11.76	6.53[4]
C-Condition	4	19.45	3.34[4]
A X B	1	0.56	[3]
A X C	4	3.25	1.90
B X C	4	4.99	2.77[4]
A X B X C	4	1.14	1.24
D-Subject (within C)	195	5.83	-
A X D	195	1.71	-
B X D	195	1.80	-
A X B X D	195	0.92	-

[1] Significant at the .1% level.
[2] Significant at the 1% level.
[3] F-ratio less than 1.00.
[4] Significant at the 5% level.

Testing of effects:
A tested against A X D.
A X C tested against A X D.
B tested against B X D
B X C tested against B X D.
C tested against D.
A X B X C tested against A X B X D.
A X B tested against A X B X D

replications, we do not believe that the large difference with milk constitutes a serious problem. The fact that the level of rating varied among experimental conditions in the first and third replications does not in itself mean very much. This could be caused by the effects of the other samples evaluated with A and B or by the psychological effects on the judges of having to rate varying numbers of samples.

The crucial source of variation is the interaction between experimental condition and specific samples. In none of the three analyses of variance involving only Samples A and B was this source significant. In the analyses of variance of the individual conditions, in only one condition of the soup and gravy base replication and in two conditions of the syrup replication was the interaction between sample and half significant. In each of these cases, the rank orders of preference in the two halves were identical. If anything, the ratings were spread out more in the second half than in the first half.

Thus, there is no evidence that increasing the number of samples up to 12 would lessen the relative differences in ratings among food samples. If there is some upper limit, we have not attained it. These negative results are meaningful and important since they give one confidence in conducting taste tests involving a greater number of samples than the typical three, four or five. However, we do not know the effect of lengthened tests upon the panel population.

Perhaps most participants would not object to an occasional ten or twelve sample test, but if such tests were to become common, then some might be induced to withdraw. Perhaps some people especially welcome frequent or short tests as breaks from their everyday activities, and might dislike less frequent and longer ones. Even so, an occasional longer test might productively be used. There is no contrary experimental evidence.

The data suggest several other problems worthy of study. A few significant interactions between sample and half were noted. It would seem that some samples are affected by certain others with which they are served, such that in the second half they drop disproportionately.

This phenomenon, which did not always appear, might be a manifestation of contrast and convergence effects (Kamenetzky, 1959). Certain samples of a product might achieve a fairly high average rating the first time, but drop when tested again, because some judges do not become aware of these deficiencies until they have had intervening experience with good quality products.

Table 7 - Summary of analysis of variance for each experimental condition (N = 40 each condition)

Conditions	Source of variation					
	Sample		Experimental half		Sample-half interaction	
	df	Signif.	df	Signif.	df	Signif.
Soup & gravy base [1]						
2-Sample	1	ns	1	5%	1	ns
3-Sample	2	0.1%	1	0.1%	2	5%
4-Sample	3	0.1%	1	0.1%	3	ns
5-Sample	4	0.1%	1	0.1%	4	ns
6-Sample	4	0.1%	1	0.1%	5	ns
Milk						
2-Sample	1	0.1%	1	ns	1	ns
3-Sample	2	0.1%	1	ns	2	ns
4-Sample	3	0.1%	1	ns	3	ns
5-Sample	4	0.1%	1	ns	4	ns
6-Sample	4	0.1%	1	1%	5	ns
Syrup						
2-Sample	1	ns	1	5%	1	ns
3-Sample	2	0.1%	1	1%	2	ns
4-Sample	3	0.1%	1	1%	3	0.1%
5-Sample	4	0.1%	1	0.1%	4	1%
6-Sample	4	0.1%	1	1%	5	ns

[1] *N* = 36 in 6-Sample condition.

Testing of effects:
Sample against sample-subject interaction.
Half against half-subject interaction.
Sample-half against sample-half-subject interaction.

If some samples are disproportionately affected when repeatedly evaluated by the same person, then one would hypothesize that they would be more subject to monotony effects if they became standard items of issue than samples which showed only the normal loss in preference in the extended test situation. The reason is that the more often a susceptible food is served, the greater the opportunity for deficiencies to be noticed. For example, if firmer evidence were available that preferences for syrup Samples C and E are nearly equal (see Table 4, 5-sample condition). one might hypothesize that preference for the former would decline more sharply than preference for the latter, assuming an equal rate of use.

NOTE: The research reported here was performed by Peryam and Kroll Research Corporation under Contract No. DAI 9-129-AMC-734(N), Pioneering Research Laboratory, United States Army Natick Laboratories. The findings are not to be construed as an official Department of the Army position unless so designated by other authorized documents. The authors gratefully acknowledge the assistance and cooperation extended by Mr. Joel L. Sidel, Project Officer and Acting Head, Acceptance Group, under whose direction all taste tests were performed.

REFERENCES

1. **Bradley, J. E., Walliker, C. T., and Peryam, D. R.** 1954. Influence of continued testing on preference ratings. In "Food Acceptance Testing Methodology," eds. Peryam, D.R., Pilgrim, F.J., and Peterson, MS. National Academy of Sciences - National Research Council.
2. **Brandt, D. A. and Hutchinson, E. P.** 1956. Retention of test sensitivity. *Food Technol.* 10, 319-420.
3. **Girardot, N.F., Peryam. D.R. and Lockhart E.E.** 1968. Relative efficiency of paired comparisons and rank order in preference discrimination. Presented at the 28th Annual Meeting of the Institute of Food Technologists in Philadelphia.
4. **Kamenetzky, J.** 1959. Contrast and convergence effects in the rating of foods. *J. Applied Psychol.* 43, 47-52.
5. **Laue, E., Zlobik, E. T. and Ishler, N. H.** 1954. Reliability of taste testing and consumer testing methods. *Food Technol.* 10, 201-203
6. **Mitchell, J. W.** 1956. Duration of sensitivity in trio taste testing. *Food Technol.* 10, 201-203.
7. **Pfaffman, C., Scholsberg, H. and Cornsweet, J.** 1954. Variables affecting different tests. In "Food Acceptance Testing Methodology." eds. Peryam, D.R., Pilgrim, F.J., and Peterson, M.S., National Academy of Sciences National Research Council.
8. **Sather, L.A. and Calvin, L.D.** 1960. The effect of number of judgments in a test on flavor evaluations for preference. *Food Technol.* 14, 613—615. received 7/5/68; revised 12/20/68; accepted 11/10/69.

III

Edited Letter to a P&K Client re: The Interpretive Value of the 9-Point Hedonic Scale

Dear ****:

You talked with Bev while I was on vacation, seeking information about possible interpretations of hedonic scale ratings. I have been elected to transmit our words of wisdom.

As you will understand, such interpretation is far from an exact discipline. So many things, besides product quality per se, can affect the levels of rating that one cannot anticipate them all. There can be a considerable range of error. But we have acquired much useful knowledge based upon our long experience in testing many types of food and food-related products. We often feel that we know what the results on a given product mean in a broad sense.

We really don't like to interpret results and evaluate products on the basis of the absolute levels of rating. As you know, it is much better to test against a standard that is included in the same test. In this way you have a specific benchmark and can interpret in relation to it. But upon occasion we stick our necks out. For example, here is a direct quote from one of our reports, which we called a rule-of-thumb interpretation. "Any average over 7.00 is good, one over 7.50 is very good, and ratings even approaching 8.00 are beyond expectation." Note that here we were dealing only with the upper range of the scale. The purpose was the attempt to put into perspective high averages obtained for certain candy products. We felt perfectly safe in predicting very good market acceptance on this basis, with the products rating over 7.50. Again, using this experience-based criterion, we recommended abandonment of a candy project where the hopeful candidates rated below 6.50. The recommendation was accepted, and we are sure we did the client a favor to stop him from further wheel-spinning. These are a couple of success stories. One cannot always be as confident.

The following scheme is a distillation of our accumulated wisdom. It's the guide which we use when we have to speculate about the probable general meaning of a result.

Average	Interpretation
8.00	Exceptional
7.50	Very Good
7.00	Good
6.50	Run-of-the-mill
6.00	On the low side
5.50	Suspect
5.00	Very Suspect
4.50 & below	Rejected as unacceptable

One thing that should be emphasized is that any interpretation should depend upon the product, at least to some degree. Some products, or product classes, are expected to achieve high averages. Examples are candy bars, ice cream, malted milks and confections in general. A candy which rated down around the 6.50 level would be a poor prospect. On the other hand, the generally expected range for ordinary, staple foods is about 6.25-7.25. More to the point, there is reason to suspect that beverage alcohol products tend to score in the lower range. Part of this is an impression derived from beer studies done many years ago. The data are not available, but I do recall noticing that ratings as high as 6.00 were unusual. Again, I can refer to the old Seagram days when we were experimenting with preference scales. There was no 9-point hedonic scale in those days, but I remember being surprised that most averages were below the scale midpoint. Other than these generalities, we can cite data from two tests on white wines and a cream liqueur study we did last winter. The averages were:

White Wines	White Wines	Cream Liqueur	
200	1500	50	
5.65	5.57	Irish	7.24
5.42	5.34	Brandy	6.08
5.39	5.02	Amaretto	6.20
5.20	5.02	Strawberry	5.62
5.15	4.07	Peach	5.04

Note that according to our generalized interpretation scheme only one of these products (Irish, 7.24) would be in the "good" range. The best of the rest would be on the low side, ranging down into the "suspect" range. One really needs a larger, more varied data base, but I suspect we would find that, in general, beverage alcohols should not aspire to such high marks as many other types of products.

Now we have told you all our secrets. The interpretive scheme is for your own use, and not for general issue. If it should be ascribed to us, we probably would deny it. We don't want to be hedged in.

Best regards,

David R. Peryam

P&K research

www.pk-research.com
info@pk-research.com
Ph: 800-747-5522

Chicago
6323 N. Avondale Ave.
Suite 211
Chicago, Illinois 60631
Fx: 773-774-7956

Dallas
3033 W. Parker Road
Suite 217
Plano, Texas 75023
Fx: 972-769-1172

Los Angeles
2435 N. Grand Ave.
Santa Ana, California 92705
Fx: 714-543-6644

New York
1025 Westchester Ave.
Suite 100
White Plains, New York 10604
Fx: 914-220-0177

Peryam & Kroll Research Corporation