CORRELATING SENSORY OBJECTIVE MEASUREMENTS—

New Methods for Answering Old Problems

Powers/Moskowitz

STP 594

AMERICAN SOCIETY FOR TESTING AND MATERIALS

CORRELATING SENSORY OBJECTIVE MEASUREMENTS— NEW METHODS FOR ANSWERING OLD PROBLEMS

A symposium
sponsored by ASTM
Committee E-18 on Sensory Evaluation
of Materials and Products
AMERICAN SOCIETY FOR
TESTING AND MATERIALS
Philadelpia, Pa., 11-12 Nov. 1974

ASTM SPECIAL TECHNICAL PUBLICATION 594
J. J. Powers and H. R. Moskowitz, editors

04-594000-36

AMERICAN SOCIETY FOR TESTING AND MATERIALS
1916 Race Street, Philadelphia, Pa. 19103

Library of Congress Catalog Card Number: 75-26426
ISBN 0-8031-0312-3

NOTE

The Society is not responsible, as a body,
for the statements and opinions
advanced in this publication.

ISBN 0-8031-0312-3

Printed in Ann Arbor, Mich.
February 1976
Second Printing, Philadelphia, Pa.
March 1985

Foreword

The symposium on Correlating Sensory Objective Measurements–New Methods for Answering Old Problems was held at ASTM in Philadelphia, Pa., 11-12 Nov. 1974. Committee E-18 on Sensory Evaluation of Materials and Products sponsored the symposium. J. J. Powers, University of Georgia College of Agriculture, presided as a symposium chairman, and H. R. Moskowitz, U.S. Army Natick Laboratories, served as symposium co-chairman.

Related ASTM Publications

Compilation of Odor and Taste Threshold Values Data, DS 48 (1973), 05-048000-36

Sensory Evaluation of Appearance of Materials, STP 545 (1973), 04-545000-36

Manual on Sensory Testing Methods, STP 434 (1968), 04-434000-36

A Note of Appreciation to Reviewers

This publication is made possible by the authors and, also, the unheralded efforts of the reviewers. This body of technical experts whose dedication, sacrifice of time and effort, and collective wisdom in reviewing the papers must be acknowledged. The quality level of ASTM publications is a direct function of their respected opinions. On behalf of ASTM we acknowledge their contribution with appreciation.

ASTM Committee on Publications

Editorial Staff

Contents

Introduction

Committee E-18 on Sensory Evaluation of Materials and Products of the American Society for Testing and Materials organized this symposium on sensory-objective correlations for three reasons. One was to review developments in the state of the art since 1967. At the Seventieth Annual Meeting of ASTM in 1967, Committee E-18 presented a program on sensory examination of materials and means of correlating sensory properties with chemical or physical attributes. Substantial changes in applied methods have occurred since that time. Not only does this symposium describe advances in methods, but in most of the presentations new applications and new data are used to illustrate methods.

The second objective of Committee E-18 was to stimulate wider use of multivariate methods of analysis. To nonstatisticians, the methods often look so formidable as to deter some from using them. The program has been organized so that the processes followed in bringing a particular problem to a successful conclusion will be described step by step in the hope that the basic elements will be clear and the would-be user will have acquired sufficient information to know (*a*) the kinds of questions one should be asking oneself and (*b*) where to get additional information. The list of references for each presentation should be particularly helpful in that respect.

The third objective comes out of the composition of Committee E-18 itself. It is composed of experts in sensory analysis, psychologists, food scientists, chemists, and scientists or engineers from other disciplines. In our committee work, we all profit from the information acquired from others working in a different field. In the same manner, we believe that a symposium with individuals of as diverse backgrounds as exist among the speakers and those attending should be to the benefit of all in pointing up pitfalls and in exchanging hard-earned experiences.

Ideally, for the setting of specifications, one would like to have a pure objective method. Results obtained by one analyst should then be identical, or nearly so, with those of another. Sensory analysis is somewhat different. The methods themselves are objective, but a certain amount of subjectivity enters into the results because of the nature of human beings, they being the instrument of analysis. When one is trying to determine whether specimens differ, the results are usually quite objective because the panelist's preferences need not enter into the matter. When one wishes to ascertain acceptability or

preference, then the results are naturally quite subjective. Invariably, acceptability depends upon an integrated response to odor, taste, feel, sight, and noise. We all do not attach the same weight to each one of these senses in making a decision. In fact, we cannot, for the acuity of any one individual varies according to the particular sense involved. One person may be oblivious to a fault in texture that another might consider critical, whereas the reverse may be true for flavor. This is one of the reasons there is a need for wider use of multivariate methods. Not only do correlations need to be established between sensory and objective tests, but also the importance of one sense response as compared with the others may need to be assessed. This is particularly so in consumer trials where the things experts think matter may not really be the things that determine consumer choice.

Errors of measurement exist both on the sensory and the objective side. The sensory error is generally greater because we all differ in sensitivity, prior cultural and training experience (the last applies to the analyst on the objective side, too), day-to-day equability, and the fact that each of us weights sense stimuli differently.

Notwithstanding the difficulties, the use of objective methods to supplement sensory methods has much to commend itself. Sensory panels are expensive and not always available. From the panelist's viewpoint, sometimes so much testing can become tedious. Through the proper use of objective tests, they can sometimes replace sensory examination. Examples of this are in the setting of purchasing specifications or the quality of finish goods. Resort may need to be made to panels only when close decisions are involved or when one needs to verify objective tests previously chosen, as when reformulation occurs. Another advantage of objective methods is in providing a reference against which panel results may be compared, for example, when a company must maintain panels in geographical locations far apart or to provide comparability of results when panel composition changes. Other benefits and limitations are pointed up in the various addresses themselves.

This symposium was organized so that the papers would be given chiefly by "practitioners" who would be speaking from their experience in devising or adapting methods for the solution of sensory-objective problems. The opening paper is, in essence, a "case study." It was designed to point up a typical problem, the kinds of information which had to be sought and the methods ultimately followed to arrive at a conclusion. The succeeding papers deal with various aspects of sensory or objective analyses themselves and especially with means of deriving useful correlations between the two types of measurements. Considerations of experimental design are also discussed.

There are many sensory and objective problems which still need to be solved if the two methods are to fully complement each other in the establishment of specifications and for other purposes. Committee E-18 hopes that the symposium stimulates others to put their own creative thoughts and efforts into the development of new or refined methods and into applications

in their fields of endeavor. In that way, all who need to use multivariate methods and correlation techniques will have an expanded storehouse of knowledge from which to draw.

J. J. Powers

Professor, Department of Food Science, The University of Georgia College of Agriculture, Athens, Ga. 30602; symposium chairman.

Andrew Dravnieks[1]

Approaches to Subjective/ Objective Correlations in Flavor

REFERENCE: Dravnieks, Andrew, "Approaches to Subjective/Objective Correlations in Flavor," *Correlating Sensory Objective Measurements–New Methods for Answering Old Problems, ASTM STP 594,* American Society for Testing and Materials, 1976, pp. 5–25.

ABSTRACT: Flavors are combinations of many subjective (sensory) properties, each of which depends in a complex way on the objective compositional characteristics of a product. A systematic approach to the design, control, and modification of flavor requires uncovering the functions that relate the various sensory properties to the analytical parameters, including the composition of headspace vapors. Available low-cost computerized programs are easily and routinely applicable to the search for such complex subjective/objective correlations. Allowances must be made in the data treatment for nonlinear correlations and interactions between the parameters. This approach is illustrated by a study of beer flavor.

KEY WORDS: sensory mechanisms, odors, taste, flavor, sensory evaluation, chromatographic analysis, data transformation, computer applications, subjective-objective correlations, beverages, beer

Odor, taste, mechanical mouth feel, and appearance of a food or beverage product combine to generate the sensation of flavor. These are the principal sensory dimensions of the product, and each can be further subdivided into additional dimensions. Thus, odor dimensions include intensity, pleasantness or unpleasantness, and many quality dimensions such as the degrees of floral, smoky, musty, etc., qualities.

Evaluation of the sensory dimensions of the flavor of a product by various types of human test subject panels may be conducted by a variety of techniques [*1,2,3*].[2] It results in a multidimensional (many characteristics considered) sensory characterization of each sample. The rating scales usually represent the degree of applicability of each sensory characteristic to the sample.

[1]Technical director, Odor Sciences Center, IIT Research Institute, Chicago, Ill. 60616.
[2]The italic numbers in brackets refer to the list of references appended to this paper.

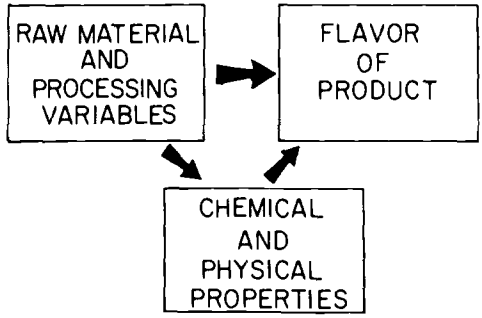

FIG. 1—*Two approaches to guidance in modifying the product.*

In product development, improvement, and manufacturing, it becomes important to know which changes in raw materials and processing variables are needed to achieve some desired combination of sensory characteristics of the product or to remedy some deficiency. Such know-how can be accumulated by experience, aided by intuition and statistical evaluation of correlations between the properties of raw materials, compositions, and process variables on one hand and the characteristics of the product on the other. This route is represented by the upper path in Fig. 1.

In the alternate route, represented by the lower path in Fig. 1, one may document the properties of the product in objective terms (composition, pH, viscosity, etc.), learn how these properties influence flavor, and then use the vast resources of chemical and manufacturing knowledge to modify the *objective* properties so as to shift the flavor dimensions to reach the desirable values. From technological knowledge, it is more straightforward to modify a process or raw materials to decrease (or increase) the content of certain substances than to find by panel tests whether some process modification would change a certain flavor property.

To effectively pursue this alternate route, it is necessary to establish the nature, direction, and degree of dependence of some sensory dimension, for example, "fruitiness," on the numerical values of various objectively measurable properties. Only rarely is the correlation relatively clear cut to one property only. Thus, the saltiness of a glass of saline water will depend mainly on the concentration of salt dissolved in water.

In most cases, the dependence is so complex that significant chances exist of either missing useful correlations or arriving at hasty decisions, unless data analysis goes beyond simple inspection and hand plotting.

This is exactly the area where present computer technology may offer invaluable help. Modern computers can conveniently and inexpensively probe for extremely involved correlations. Their use requires only elementary understanding on how to prepare data and which computer programs to use.

The objective of this paper is to summarize the main considerations involved in a search for such complex correlations. Data on beer are used as the vehicle

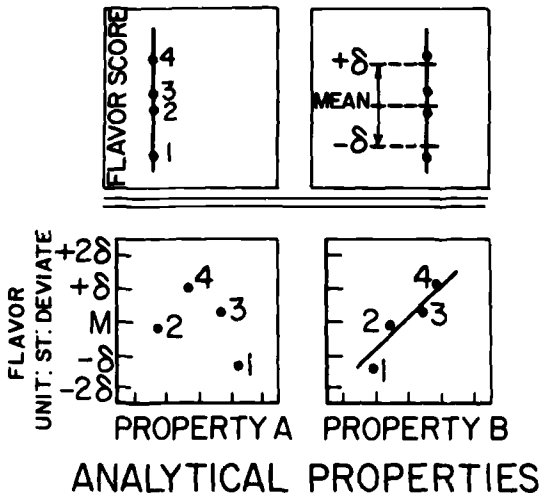

FIG. 2—*Linear correlation between a flavor dimension and objective Property B.*

for illustrating this search process. The problem may be defined as exploring the feasibility of characterizing beer flavor from analytical and gas chromatographic headspace analysis data.

Primer on Types of Correlations

A few simple geometrical sketches may clarify the nature of the approaches in the search for subjective/objective correlations.

Continuous Property

A property such as a score of flavor goodness may have any value permitted by the selected scale. Thus, if the flavor goodness is rated on a 0 to 100 scale, any value in between is possible. Figure 2, upper left, represents raw data for four samples. The upper right sketch is an elementary analysis of such data in terms of a mean value and standard deviation. Not much more can be derived from such data only.

Now, suppose that two analytical properties, A and B, are also known for each of the four samples. The two bottom sketches are attempted plots of the flavor goodness versus Property A and versus Property B. They suggest that flavor is linearly related to Property B, but not to Property A.

However, other types of correlations may occur (see Fig. 3). The left sketch indicates that a relation, although more complex than versus Property B, may exist also versus Property A; for some value of Property A, the flavor is best, and it deteriorates both with an increase or decrease in Property A.

The two other sketches in Fig. 3 indicate still another possibility. Here flavor obviously does not relate well, in a linear form, to still another

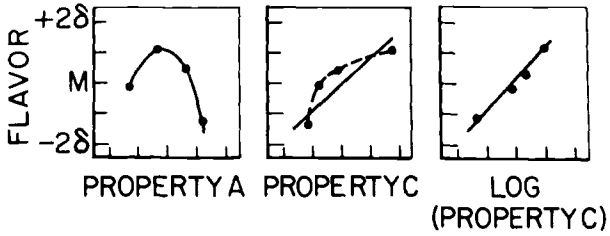

FIG. 3–*Parabolic and logarithmic correlations.*

analytical property, C. However, the curve that connects the points straightens out if the *x*-axis is replotted as logarithm *C*.

Since various properties interact, more complex relations may occur. For instance, the position of maximum in the left sketch of Fig. 3 (for example, the optimum amount of spices in soup) may shift as another property, D, (for example, concentration of salt) is varied (Fig. 4). Simpler interactions may also be observed, for example, the position of the plot in Fig. 2 lower right may shift as Property D is varied.

Improvements in correlations may occur as additional terms are considered. Thus, Case B, Fig. 2, may result in a better correlation if the four samples possess another property, for example, E, which has values such that if multiplied for each sample by the same numerical factor, the resulting increments can be used to correct the positions of points which then align in a better straight line.

Class Property

Here samples are classified in groups, for example, "bad" and "good," or different subtypes of the same product. The task is to find which combination of analytical properties results in a correct classification.

Figure 5 illustrates the principle. Suppose that Samples II and III are good, and I is bad. Suppose that a gas-chromatographic (GC) analysis characterized

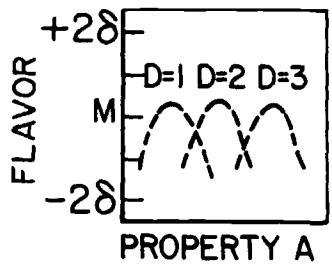

FIG. 4–*Dependence of optimum value of Property A on values of Property D.*

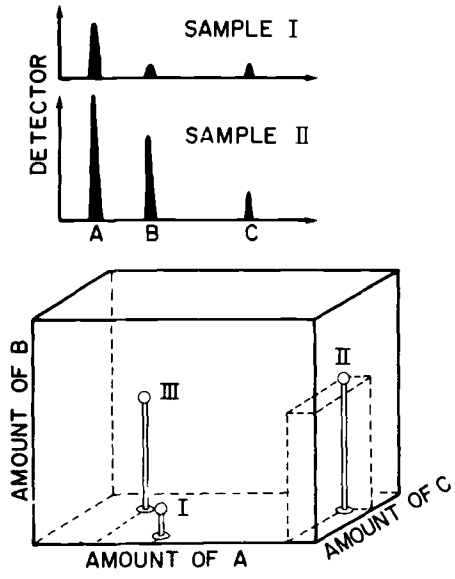

FIG. 5—*Classification in multidimensional space.*

each sample in terms of the concentration of three components. This is shown for two of the samples by idealized gas chromatograms at the top of Fig. 5. Areas under the three GC peaks represent the concentration of the components.

If the concentrations of each of the three components are considered as separate coordinates, then each sample can be represented by a single point in a three-dimensional space, as in Fig. 5. Any number of samples can be so represented. If more GC or analytical properties are measured, the space must be extended to more than three dimensions. Although this higher space cannot be plotted *in toto*, it still is a valid mathematical construction.

The relations between properties and classifications are explored by a mathematical process which in essence reduces to the following graphic concept. Inspect Fig. 5, bottom. Imagine yourself moving around the depicted cube and looking at the arrangement of the three points. Try to find a direction looking from which the two good points Samples II and III superimpose or appear to be as close as possible ("cloud" of good samples) while the bad point Sample I is as far from this cloud of good samples as possible. Such a position would be on the left side, about half way up on the side of the cube. From this position, a flat projection map could be drawn on which Samples II and III would be together, while Sample I will be far apart.

With, for example, a seven-dimensional space, the initial representation of many samples will require all seven dimensions. The task for correlations of classification to analytical properties consists of (*a*) probing whether fewer properties would suffice, (*b*) if so, which ones; and (*c*) usually, which is the

simplest two-dimensional representation which best separates the clouds representing the specific classes. It is important to realize that the number of samples must be much larger (at least twice) than the number of dimensions.

Programs

The process of probing for various types of correlations that might exist is easily accomplished by available computer programs. They rapidly test for the goodness of hypothetical correlations and produce the necessary plots.

There is a great variety of programs for probing the existence of correlations between variables. It is not the intent of this paper to discuss tailoring specific programs to specific needs, but rather illustrate utilization of only one of the programs, Biomed 02R.

When the sensory property is of a continuous type, mentioned previously, stepwise regression analysis is a suitable tool. When the samples are classified into sensory classes, stepwise discriminant analysis is suitable (it searches for properties which best discriminate between classes).

In this paper, only the regression approach will be illustrated. Examples of application of stepwise discriminant analysis can be found elsewhere [*4-7*].

Caution

The outlined methods result in indications of possible correlations. Additional steps, that cannot yet be illustrated by the following examples on beer, are necessary to establish the true significant correlations.

The simplest next step is to test the predictive merits of the obtained correlations. This is done by applying the developed correlation equations to another set of samples not used in the initial search for correlations. If the resulting predicted values of (for example) flavor goodness approximate the panel-given values (at some statistically significant levels), then the first test on the predictive merits of correlations is passed.

Confirmation of the true significance of specific analytical properties in their influence on the flavor is the final and laborious phase. The need for such a phase stems from possible inherent correlations among the analytical properties. Thus, rancidity of vegetable oils may correlate with the pentane content in the oil vapors, but since pentane *per se* does not cause the rancid odor, this is an indirect correlation. Artificial addition or removal of pentane to oil would not influence rancidity.

A component cannot be considered as truly correlated to flavor unless it can be established that its addition or removal indeed changes the flavor in the expected direction and to the expected extent. If this does not occur, it is likely that some other components, usually co-occurring and perhaps reaction-mechanisms-linked, are the real flavor-influencing substances. In such cases, the component found to correlate with flavor may serve as an "indicator" of the presence of others. This, however, would not apply to such changes in materials and processes that would change the concentrations of the truly

TABLE 1—*Some data on 31 beer samples.*

Sample	Flavor Profile Percentages	Dimensions		Content of F & A 26 GC Variable, Arbitrary Units
		Fruity Floral	Bitter	
1	98	4.9	5.0	40
2	88	4.8	4.7	(0)
3	60	4.6	5.0	26
4	93	4.8	5.0	(0)
5	64	4.6	5.0	17
6	53	4.8	4.9	46
7	40	4.8	5.0	5
8	86	4.8	5.1	36
9	48	5.0	5.1	(0)
10	55	4.6	4.8	(0)
11	81	4.5	4.8	38
12	20	4.5	4.5	59
13	35	4.6	4.7	(0)
14	44	4.6	4.6	22
15	88	4.9	5.2	16
16	24	4.6	4.8	39
17	91	5.3	5.5	19
18	60	4.7	4.7	30
19	94	4.6	4.6	30
20	78	4.0	4.1	23
21	81	4.3	4.8	61
22	95	4.7	4.7	24
23	40	4.6	4.4	49
24	57	4.5	4.5	40
25	78	4.7	4.9	33
26	60	4.4	4.2	42
27	81	4.3	4.5	43
28	71	4.6	4.7	51
29	64	4.4	4.7	22
30	53	4.3	4.5	52
31	78	4.3	4.7	92

influential components without changing the concentration of the indicator component.

Data

Subjective Data

Sensory evaluations on 31 different samples of beer were available that contained panel data on several dozen of sensory flavor, taste, and the odor dimensions. The rating was in terms of scores. For the purpose of the present discussion, two specific sensory dimensions, "fruity floral" and "bitter," and one more general, "profile" were selected. The profile refers to the overall flavor quality of beer and its values in the set of samples here range from 35 to 98, (scale percentage values). Bitter is principally a taste variable while fruity floral is principally an odor variable. The first three data columns, Table 1, list the raw sensory data for the 31 samples.

Objective Data

These subdivide in two groups. One group, termed here analytical data, includes routine measurements of apparent extract, real extract, percent alcohol, real degree of fermentation, pH, titrable acid, formol nitrogen, calcium, isohumulone, oxalate, free and total sulfur dioxide (SO_2,) carbon dioxide (CO_2), and air, as well as the content of several higher alcohols and ethyl acetate and isoamylacetate. Space does not permit reproduction of the entire set of data.

The other group of objective data resulted from headspace analysis. Volatile odor sensation causing substances are primarily responsible for the aroma component of the flavor and do so because their vapors can reach the olfactory receptors in the nose. The concentrations of such vapor substances reaching the nose should therefore be more directly related to their concentration in vapors *above* the beer than to the concentrations *in* the beer. The concentration in the vapor above the beer depends not only on the vapor pressure of each substance but also on the physicochemical interactions between the substances and water. Thus, two substances, one polar, another nonpolar, with the same vapor pressure in the pure state and dissolved at the same concentration in water, will vastly differ in their concentration in vapors above the water; the headspace will be relatively much richer in the nonpolar substance than in the polar.

Headspace vapors were collected by passing helium over the beer surface in a horizontal glass tube and absorbing them in high-surface area organic polymer phase. This enrichment process and the procedure of the sample transfer to a gas chromatograph have been described elsewhere [*8*]. It permits escalating the GC flame ionization detector sensitivity to a level where substances present in the headspace at odor-causing concentrations will produce a definite GC peak, even if their odor threshold is very low.

It has become increasingly common to combine GC headspace analysis with a simple sensory evaluation of odors of gas-chromatographically separated components at their emergence from the GC column. For this purpose, the GC column effluent is split with, for example, 50 percent flowing to a flame ionization detector (FID) and the rest exiting through a sniffing nozzle [*9-11*]. The analyst notes the odors of the separated components. The result of such a sensory assay technique is an "odorogram" where the amounts of the odorants are quantitatively expressed by areas under GC peaks, while their odors are estimated by the nose. In cases where a GC peak corresponds to two unresolved components, a small amount of odorous substance and a larger amount of a non-odorous substance, the GC area may misrepresent the concentration of the odorant.

Table 2 illustrates a computer print-out of a gas chromatogram of beer headspace volatiles. Each peak's retention time, the area under the peak and the odor descriptor were entered on a separate punched card. A computer program organizes these cards in the table shown; also, it calculates and prints

TABLE 2—*Typical GC/sensory assay: data on a beer sample; Carbowax 20M column; sample designation: 13; total sample size (integrator units): 17861.*

Peak	Retention Time, min	Peak Area, Integrator Units	Area % of Total	Odor Note
1	5.40	161	0.90	yeasty
2	6.00	161	0.90	cheesy, sour
3	7.58	69	0.39	caramel
4	7.95	67	0.38	chocolate
5	8.66	51	0.29	citrus
6	8.85	1597	8.94	acidic
7	9.23	1065	5.96	vinegar
8	10.20	8396	47.01	yeasty, malty
9	11.00	...	...	ethanol
10	16.95	266	1.49	pleasant, estery
11	18.15	211	1.18	grainy
12	19.13	246	1.38	musty
13	19.31	543	3.04	caramel
14	20.10	292	1.63	paint-like
15	22.73	32	0.18	grainy
16	23.10	1	0.01	sour, unpleasant
17	25.43	29	0.16	floral, very pleasant
18	25.76	901	5.05	caramel
19	26.03	3727	20.87	acetic acid-like
20	28.50	3	0.02	floral
21	36.45	1	0.00	floral
22	38.55	0.3	0.00	musty
23	40.68	18	0.10	estery, floral
24	44.25	3	0.02	grainy
25	44.85	3	0.02	cheesy
26	45.75	1	0.01	floral
27	47.93	0.3	0.00	perfumery
28	48.90	2	0.01	turpentine-like
29	56.85	1	0.00	hoppy
30	59.85	0.5	0.00	floral
31	62.03	0.4	0.00	floral
32	64.95	2	0.01	rose-like
33	65.48	3	0.01	hoppy, rose-like
34	73.80	2	0.01	hoppy
35	78.75	0.5	0.00	strong, hoppy
36	80.85	1	0.01	hoppy, oily
37	83.10	0.7	0.00	beery
38	89.40	0.4	0.00	hoppy
39	97.95	0.3	0.00	hoppy

out the area of each peak in terms of percent of the sum of areas of all peaks. This column indicates the approximate relative participation of each peak in the total content of organics (peak of ethanol is ignored in the beer).

GC Data Preparation

No problems exist in representing each sample by an appropriate value of pH or some other definite analytical property. For GC data, difficulties arise

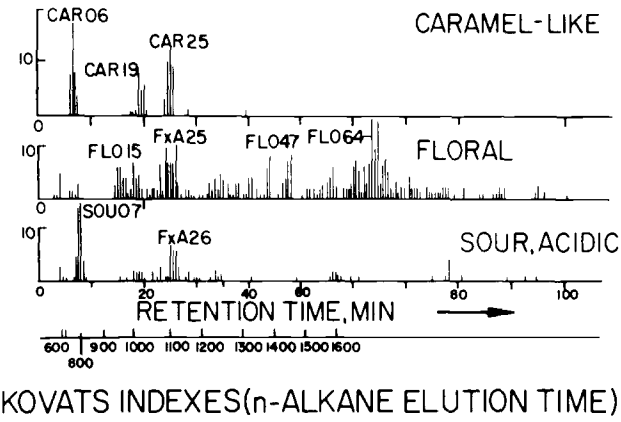

FIG. 6–*Histogram of the frequency of occurrence of components with specific odors at various positions in gas-chromatographic pattern of beer headspace volatiles.*

since the GC retention time for the same substance slightly shifts from analysis to analysis, and the chemical identity of the substance corresponding to a peak frequently remains in doubt. A complete mass-spectrometric identification of each peak-represented component in each sample is an impossibility for any practical purpose.

One approach, utilized in the present study, is to reorganize the GC data so that each sample could be described by content of what appears to be frequently occurring odorants. The first step is to prepare a histogram of odor-bearing peak occurrences along the retention time axis, separately for each type of odor. This can be done manually, but is more conveniently possible using a specially written computer program. Figure 6 shows the resulting histograms for several odor types.

An inspection of such histograms indicates that peaks, for example, with a floral or floral-acidic odor character occur predominantly at certain retention time zones. Thus, each gas chromatogram (each sample) can be expressed in terms of its content (GC peak area size) of a floral-acidic component, F & A 26, occurring in the vicinity of the 26th minute. From GC calibration with known compounds, margins of ±1.5 min in retention time (at longer retention times) with respect to the frequency maximum are allowed; a floral peak at 27 min is assigned to belong to F & A 26 variable. This may be a compromise in which much information is sacrificed.

By this method, 27 GC/sensory variables were defined, and each sample was characterized in terms of the content (area under the peak) of each of these variables. These variables are listed in Table 3; numbers at the end of three-letter code are the mean retention times, in minutes, for positions where the particular peak associated with the particular odor most frequently occurred.

This procedure is simply a first rough organization of the GC/sensory assay data. It does not include provisions for incomplete resolution of peaks, for an

TABLE 3—*Odor component variables of beer head space derived from histograms such as in Fig. 6.*

Sample	Odor Character	Retention Time min	Code Designation	Number of Beer Samples in Which This Component was Observed
1	yeasty	5	YEA 05	31
2	cheesy	5	CHS 05	31
3	sour, acidic	7	SOU 07	31
4	medicinal, phenolic	7	MED 07	9
5	medicinal, phenolic, burnt	23	MED 23	18
6	caramel	6	CAR 06	31
7	caramel, malty	19	CAR 19	27
8	caramel, malty	25	CAR 25	30
9	malty	25	MAL 25	21
10	turpentine, paint	18	TER 18	15
11	turpentine, paint	46	TER 46	10
12	pungent, sharp, harsh	47	PUN 46	21
13	beer-like	75	BEE 75	26
14	unpleasant, repulsive	28	UNP 28	11
15	stale, oxidized	29	STA 29	7
16	stale, oxidized	50	STA 50	6
17	hoppy	75	HOP 75	31
18	hoppy	96	HOP 96	31
19	floral	15	FLO 15	26
20	floral	24	FLO 24	30
21	floral, fruity	47	FLO 47	30
22	floral, prominent roselike	64	FLO 64	30
23	floral and acidic	26	F & A 26	31
24	musty	30	MUS 30	9
25	musty	37	MUS 37	9
26	musty	79	MUS 79	7
27	oily as principal note	86	OIL 86	9

occasional larger shift of retention time, missing of an odor by the analyst, etc. No doubt, the procedure can be further improved, especially by a better GC resolution and identification of the corresponding odorants. A column in Table 1 illustrates as an example, the content of the F & A 26 component (No. 23) in the headspace of the various beer samples.

Transformations of Variables

Our task is to find the mathematical relations between the subjective flavor dimensions and the objective (analytical and GC) variables.

The simplest assumption is that the number representing a property such as the flavor profile consists of positive (improving the profile) and negative (detrimental to profile) increments contributed by those objective variables

which happen to influence the profile. This is equivalent to searching for an equation

$$(\text{profile}) = \text{constant} + a_1\ (\text{pH}) + a_2\ (\text{alcohol content}) + \ldots$$

with the appropriate properties selected by the computer, and the values of the coefficients also to be found by a computer program. The numbers representing pH, etc. can be used directly. Much of the demonstration material in this paper represents this approach known as linear multivariate correlations. Linear refers to variables taken without any further modifications.

Many reasons exist why certain operations, mathematical transformations of the variables, merit consideration. They can be divided into several categories, some already represented in Figs. 2 through 4.

Logarithmic Transforms

The intensity of odor or taste sensation is not proportional to the physical intensity (concentration) of the stimulus. For example, the odorant vapor concentration must be increased several fold before odor appears to double in its apparent intensity. The relation is expressed by a psychophysical power function [*12-14*].

$$(\text{intensity of sensation } I) = kC^n$$

or

$$\log I = \log k + n \log C$$

where

k = constant different for each different taste and odor stimuli,
C = concentration of the substance in the stimulus, and
n = exponent which may be different for different types of stimuli.

Thus, the use of the logarithm of the concentrations instead of the concentration directly may be justified. Some variables, for example, pH, are already logarithmic transforms of the underlying variable.

Sensory dimensions can also be expressed in different scales. The preceding equation applied to the magnitude of the intensity of sensation, $I = 20$ would be a sensation twice stronger than $I = 10$. Many sensory dimensions are rated in terms of category scales, such as scores. In the odor intensity category scale, consisting of, for example, 0,1,2,3,4, and 5, a rating of 4 does not correspond to an odor or taste twice stronger than of the rating 2 [*14*]. The category scale values are already approximately proportional to the logarithms of the intensity of sensation, except that resolution at the highest end usually fades out, producing a plateau effect (two odors extremely strong and both rated 5 on a 5-category scale may still significantly differ in intensity when compared directly).

Parabolic (Square) Transforms

With some objective variables, the sensory property does not increase systematically with an increase in the value of the variable, but reaches, for example, a maximum at a certain value; that is, the best flavor of soup will be reached at a certain salt concentration, and the flavor would be impaired by too much or too little salt.

To probe for the existence of maxima (or minima) in the function

$$\text{flavor dimension} = \text{function (concentration of component A)}$$

a mathematical form is used

$$I = k\ C^2 + \text{constant}$$

This represents a parabola, compare Fig. 3. Of course, the real shape of the correlation may be much more complex, but if a correlation improves when the objective variable is used in the form of its square, existence of an optimum concentration where the sensory property peaks is indicated.

Interaction Terms

A variable, for example, titrable acid in beer, may have a different influence on the flavor at different values of some other objective variables, for example, higher alcohols. Such a relation is probed for by a transform

$$\text{flavor} = k_2\ (\text{titrable acid})\ (\text{concentration of higher alcohols})$$

This is an interaction term (cross term) representing one of the simplest forms of interactions between variables (others may be ratios of the variables, etc.). Higher interaction terms may also exist, for example

$$\text{flavor} = k_3\ (\text{concentration of } A)\ (\text{concentration of } B)^2$$

Here an optimum concentration of B (for example, salt in soup) is probed for, allowing for the possibility that this optimum shifts with a change in the concentration of A (for example, spices), compare Fig. 4. This is a so-called third-order term (sum of exponents A^1B^2 is 3).

Relative Concentration

In the utilization of GC data, two approaches are possible. First, a flavor dimension may depend on the absolute concentration of some components. Second, since an aroma is recognizable over a broad range of dilutions, the ratios between the various components, or the content of some component in terms of percent of total concentration of all odorants present may be important (compare Table 2).

Utilization of Computers

The computer program can be likened to a mathematical operational symbol. In elementary mathematics, symbols such as +, $\sqrt{\quad}$, $\log_{10}$, etc., indicate simple operations of adding, taking square root, finding the logarithm of a value, etc. A computer program used for the illustration here has a code name Biomed 02R. This may be thought as an operational symbol which tells what the computer should do with the numerical data supplied. It is not necessary to understand in detail how it does it, just as it is unnecessary to understand, when using an electronic pocket calculator, how it multiplies or takes logarithms.

Elementary Remarks on the Use of Computers

Computers can receive experimental data and instructions on how to handle these and what to do with them in a variety of forms. One most commonly used utilizes punched cards. Thus, data must be first reduced to the punched card format. This operation is termed keypunching and is somewhat similar to secretarial typing, but uses a special machine in which key strokes produce holes in appropriate places in the card. Each card has space for 80 entries, each entry being either a single digit, single letter, or one of a few selected symbols.

The usual preparation of data and operating instructions for keypunching is by listing data on special coding sheets. These contain rows of 80 spaces for entering by hand the information in the exact form in which it will appear in punched cards. Once the information is entered, keypunching can be done without any understanding of the meaning of the information. Keypunching services are available within many organizations and also commercially from companies specializing in data processing services.

Instructions on what the computer must do with the information in the cards are also supplied in punched card form. Preparing such instructions requires specialized knowledge and is termed programming. For programs discussed in this paper, where complex mathematical information must be performed, all the required instructions have been already prepared by the developers of the computer program. The only remaining programming, to make the program operational, is to supply, again in punched card form, a few elementary instructions. The format which should be followed in coding these instructions on coding sheets is explained in manuals describing the specific program. The manuals are available from suppliers of computer services. These instructions list the location of the information on the experimental data cards, prescribe how to enter labels for variables, select certain options (which variables to ignore, etc.). For a researcher who would want to keep preparing such instructional code sheets by himself, a good working session with some individual familiar with the elements of programming may suffice.

Thus, becoming operational with some specific computer program is a relatively simple matter and can significantly increase the scope of findings that can be derived from voluminous experimental data.

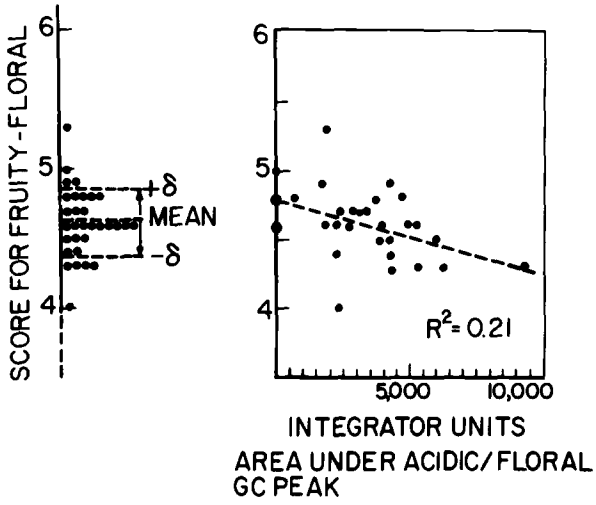

Fig. 7—*Initial correlation of floral-fruity dimension of beer to a gas-chromatographically indicated odor component.*

Essence of Correlations

We have scores of floral/fruity dimension of 31 beer samples (Table 1). These can be organized as in Fig. 7, left, and a mean value can be calculated for the entire group. Limits of one standard deviation are shown. This display has rather low predictive value; if a new beer sample is picked at will, its floral/fruity score would have more chance to be closer to the mean than far away; in two cases of three, the new sample will be within the standard deviation limits shown.

But we also have objective data on the same samples. The computer finds that among these, the area of a GC peak corresponding to a Component 23 from Table 3, correlates to the floral/fruity score better than the others. This correlation is shown at the right of Fig. 7. If we know the area of this GC peak of some new sample, we are in a somewhat better position to predict the floral/fruity score of this beer sample.

When the computer tries to improve on the correlation by adding information on an additional objective variable, in this case Component 27, Table 3, correlations improve. Figure 8 shows plots of score values calculated from the computer-found correlations versus the actual scores. Plots after taking into account 1, 2 and finally 6 variables are shown. These were obtained using the Biomed 02R computer program. The variables of importance were 4, 7, 18, 23, 26, and 27 of Table 3.

Biomed 02R, Stepwise Regression Analysis Program

This program is already contained in the service memory of many computers in use for engineering and scientific purposes and available in other

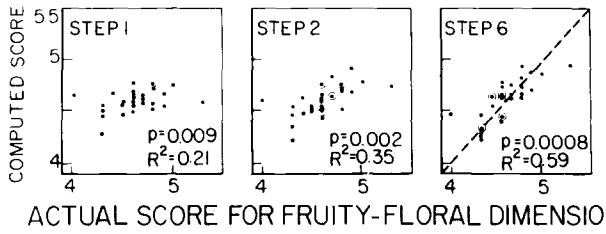

FIG. 8–*Improvement in correlation of Fig. 7 with addition of further gas-chromatographically indicated components.*

preprepared forms for putting it into the computer memory when needed. In the case when it is a part of the service memory, the computer needs: (*a*) punched code card that recalls the program from memory, to be ready to be used, (*b*) punched cards that contain experimental data, as well as some code for the sample identification, and (*c*) punched card instructions that tell the computer where (in which column) the specific experimental information can be found in the experimental data cards, as well as some general information about the problem.

The instruction cards state the following.

1. Which is the dependent variable? In one of the following examples it is the flavor profile value, since we are trying to find how it depends on the objective properties.

2. Which are the objective properties which we want to probe for possible influence on the flavor profile, and which do we wish to ignore? These are independent variables.

3. Should these properties be utilized in direct form as in the data cards or transformed to some other mathematical forms (taken as logarithm, or squared, or another derived property formed by multiplying of the values of two or more properties). The computer program has the capability to effect these transforms when asked.

4. General information is stated showing how many samples are included, how many variables, etc.

5. When to discontinue the analysis is shown; an option in the program instructions provides for economy in stopping the calculations if the statistical significance of the remaining terms sinks below some prescribed probability level.

6. What needs to be printed out in the result data is shown (correlation tables, table of actual values and the computer-predicted values, plot of certain variables versus other variables, etc.).

Some Characteristics of Stepwise Regression Program

This program searches for correlations between a "dependent" variable (for example, fruity/floral score) and many "independent," including transformed,

variables. It operates in a stepwise fashion. It first inspects each objective independent variable singly for best correlation to the dependent variable. The one which is the best results in the equation

$$(\text{sensory score}) = \text{constant} + k_1\ (\text{independent variable } X)$$

with the coefficients printed out by computer.

The program then reinspects the data and finds out which of the remaining independent variables improves the correlation best and adds a corresponding term to the equation. This second best variable is not necessarily the same that was the best second in the initial correlation table, since the initially best second may have provided very similar information as the initial first; after the above equation was generated, this initially second best may contain very little additional useful correlation information and is then ignored by the computer.

It is extremely important to remember that some correlations may be by chance only.

Thus, if there are 30 samples, 50 variables, and the computer is allowed to keep adding terms to the equation until all variables are used up, it is very likely that for each of the samples some individual feature may be included. This would result in a highly provincial solution, with "correlations" valid for only this set of data. Chances of such useless result become less if the number of samples exceeds the number of variables by more than a factor of 2. It is much more realistic to keep this ratio much higher than that. It appears that looking for more than ten correlated variables in most sensory problems is a needless exercise which simply produces "noisy" information. If useful correlations are not found after four to five variables, they are most probably not there. To plan for searching for five more generally useful variables that correlated to the dependent variable such as flavor profile probably at least 15 to 20 samples, preferably more, should be used. The computer can be instructed to stop after finding five terms (plus constant) in the correlation equation.

At stopping the calculations, the computer produces summary tables. Let us review a typical sample data result.

One of printouts gives the coefficients for the final correlation equation, with codes denoting variables.

Another printout produces a table which lists, for each sample, the actual value of the flavor profile and the value obtained when the developed equation is used. Another table also contains the difference between the experimental and the equation-predicted value. Figure 9 has been drawn by utilizing such printouts.

A plot of the calculated versus the actual experimental variable is also produced.

Finally, the printouts also contain information on the statistical significance of the equation and of the terms within the equation. One such significance

index is the multiple correlation coefficient, or rather its square, R^2 (coefficient of determination). A coefficient of one (unity) would indicate complete agreement between the actual and the equation-predicted value. However, another index, called F-ratio, permits a better estimation of the significance. Tables [*2*] exist which list critical F-ratio values for various "degrees of freedom," a statistical term related to the numbers of samples and utilized variables.

In this case, Fig. 8, last sketch, the applicable numbers of degrees of freedom are 30 (one less than the number of samples) and 6 (six variable terms in the correlation equation). The computer-calculated value of the F-ratio is 5.75. Tables [*2*] indicate that an F-ratio of 3.47 must be exceeded for $p < 0.01$ level (99 percent confidence level) of statistical significance. Since $5.75 > 3.47$, the correlation such as found could arise by chance in less than one case in a hundred. The particular variant of the computer program is equipped with a printout for p (probability of such result by chance only) by exact calculation and gives $p = 0.0008$; such a result could occur by chance in eight cases out of ten thousand, which is a rather convincing correlation. In sensory research, correlations at $p > 0.05$ are considered sufficiently significant (one chance in twenty to obtain such correlation by a statistical accident).

The printout also produces separate F-ratio values for significance of the various terms in the correlation equation. Finally, a table is produced listing the F-ratios for each of the variables *not used* in the equation; it gives an insight into the remaining information.

Quality of Predictive Correlations

Unless the correlations between sensory dimension and objective properties are real, they are of little practical use. The correlation equation should be generalizable and not provincial (fitting just this group of samples).

Determining whether the equation is generalizable can be done by attempting to apply the equation to another set of samples, those not used in the development of the equation. A provision in some computer programs permits loading into the calculation data on all samples, but using only a part of the samples for the equation development, then utilizing the equation to calculate the predicted values for the withheld samples. Statistical tests exist to estimate how well the equation worked on the additional samples.

It is evident that for considering 5 variables in the correlation, 25 to 30 samples must be analyzed; 15 to 20 could be used to develop the equation and the rest to test the equation for its competence in predicting the flavor profile value on new samples.

Examples of Correlations

Flavor Profile—This represents a combination of odor and taste dimensions. Figure 9 indicates that correlation to routine laboratory analytical data is satisfactory, also to the headspace GC analysis data. Correlations improve when both analytical (more taste related) and headspace analysis (more aroma

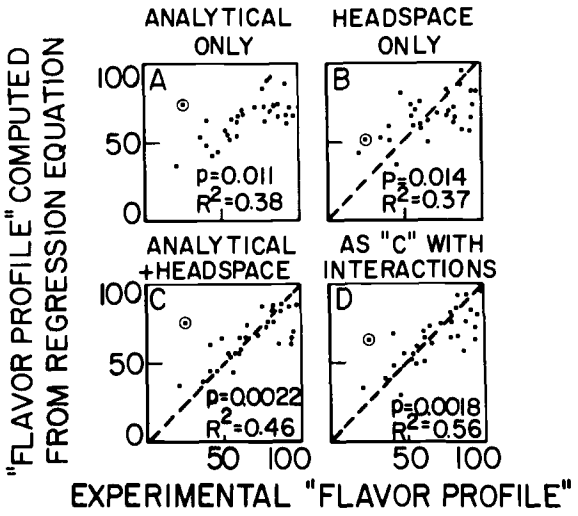

FIG. 9—*Correlations of flavor profile dimension of beer to analytical, gas-chromatographic, and combined objective properties.*

related) data are used. The correlation also further improves when cross terms are introduced to probe for interactions between the properties.

The headspace GC data here were used in the form of peak area sizes. A computer test was conducted to test if the peak areas taken in terms of percent of the sum of all peak areas are more significant than areas alone. The value of R^2 increased from 0.37 to 0.51, and only two GC variables were used both in the direct and percent area correlations, while the other four were different.

Correlation of the flavor profile to a combination of both analytical and head-space gas-chromatographic properties, with some cross terms for the analytical variables (Fig. 9, bottom right) indicated that the following properties were significant:

(*a*) headspace gas-chromatographic variables, components smelling, yeasty, malty, and terpene-like (1, 9, and 11 in Table 3); and

(*b*) analytical, percent alcohol, formol nitrogen, titrable acid, n-propanol, and isoamylalcohol contents. Although all were available for correlations as separate terms, the computer program selected to use these in the form of cross terms: (alcohol) X (formol nitrogen), (titrable acid) X (n-propanol), and (titrable acid) X (isoamylalcohol).

As an illustration, the following correlation equation resulted.

flavor profile percentage value = 2.17 + 0.044 (area yeasty peak) +
+ 0.0049 (area malty peak) -
- 1.8 (area terpeny peak)
+ 1593 (alcohol) (formol N)
- 0.144 (titrable acid) (propanol)
+ 0.052 (titrable acid) (isoamyl alcohol)

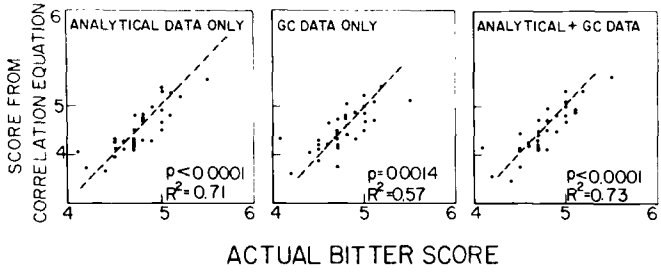

FIG. 10–*Correlation of bitter dimension of beer to objective properties of beer.*

Bitter–This is mostly a taste term; Fig. 10 shows that correlations versus headspace analysis data are poorer than versus analytical laboratory data.

The preceding examples dealt with the entire set of 31 beer samples. The merits of predictive correlations can be explored by developing correlations on a part of samples and testing them on additional samples.

Concluding Statements

Application of computerized search for a correlation between some sensory and the analytical and GC headspace data on beer indicated the emergence of useful correlations.

Thus, the problem theme which was defined as establishing the feasibility of characterizing beer flavor by correlation to its chemical and physical properties has been successfully solved. A number of additional computerized experiments with the same data remains possible, for example, testing for other forms of interactions between the variables. Testing for alternate variables that may successfully substitute for those selected by a computer may also be instructive; if two objective properties are themselves highly correlated, and both correlate well to the flavor, the computer blindly picks one which correlates slightly better, but only by chance. Testing for usefulness of the alternate objective property may result in better final correlations when other variables are added.

Practical Utilization of Findings

Obviously, long-range application of findings such as were described for beer improvement would require much more additional work. The following aspects may be identified.

1. Sensory and consumer preference studies are needed to establish the optimum combination of flavor dimensions for raising the acceptability of beer.

2. Experiments to establish if artificially modifying some objective property which was found important by computer analysis indeed will shift the expected flavor dimensions in the expected direction.

3. Investigations are needed on how to achieve the needed changes in the objective properties by raw material and process changes.

4. Market tests should be done on an improved beverage.

Acknowledgments

Gas-chromatographic study was supplied by B. K. Krotoszynski, and computer analysis was conducted by Joyce Whitfield, both with IIT Research Institute. Their contributions are gratefully acknowledged.

References

[*1*] *Basic Principles of Sensory Evaluation, ASTM STP 433*, American Society for Testing and Materials, 1968.

[*2*] *Manual on Sensory Testing Methods, ASTM STP 434*, American Society for Testing and Materials, 1968.

[*3*] Amerine, M. A., Pangborn, R. M., and Roessler, E. B., *Principles of Sensory Evaluation of Food*, Academic Press, New York, 1965.

[*4*] Powers, J. J. and Keith, E. S., *Journal of Food Science*, Vol. 33, 1968, pp. 207-213.

[*5*] Young, L. L., Bargmann, R. E., and Powers, J. J., *Journal of Food Science*, Vol. 35, 1970, pp. 219-228.

[*6*] McDaniel, H. C., *American Perfumer and Cosmetics*, Vol. 83, 1969, p. 34.

[*7*] Dravnieks, A., Reilich, H. G., Whitfield, J., and Watson, C. A., *Journal of Food Science*, Vol. 38, 1973, pp. 34-39.

[*8*] Dravnieks, A., Krotoszynski, B. K., Whitfield, J., O'Donnell, A., and Burgwald, T., *Environmental Science and Technology*, Vol. 5, 1971, p. 1220.

[*9*] Guadagni, D. C. in *Correlation of Subjective-Objective Methods in the Study of Odor and Taste, ASTM STP 440*, American Society for Testing and Materials, 1968, pp. 36-48.

[*10*] Dravnieks, A. and O'Donnell, A., *Journal of Agricultural and Food Chemistry*, Vol. 19, 1971, pp. 1049-1056.

[*11*] Dravnieks, A., Keith, L., Krotoszynski, B. K., and Shah, J., *Journal of Pharmaceutical Sciences*, Vol. 63, 1974, pp. 36-40.

[*12*] Stevens, S. S., *Psychological Review*, Vol. 64, 1957, p. 153.

[*13*] Cain, W. S., *Perception and Psychophysics*, Vol. 6, 1969, p. 349.

[*14*] Dravnieks, A., *Tappi*, Vol. 55, 1972, pp. 737-742.

Elizabeth Larmond[1]

Sensory Methods—Choices and Limitations

REFERENCE: Larmond, Elizabeth, "Sensory Methods–Choices and Limitations," Correlating Sensory Objective Measurements–New Methods for Answering Old Problems, ASTM STP 594, American Society for Testing and Materials, 1976, pp. 26–35.

ABSTRACT: Several methods of sensory evaluation have been developed, and each has its advantages and disadvantages. The experimenter must be aware of these and select the most practical and efficient method in each situation. The form of the results obtained depends on the method used and determines what further treatment can be applied to the data.

KEY WORDS: sensory mechanisms, scale (ratio), tests

Sensory evaluation is extremely valuable in the measurement of food quality since no instrument can perceive, analyze, integrate, and interpret a large number of sensations at the same time. If the human mouth could be duplicated in an instrument, we would still have to duplicate the nervous system which has the important function of receiving and translating the signals from the receptors in the mouth.

Sensory evaluation panels can be grouped into three types: (*a*) highly trained experts, (*b*) laboratory panels, and (*c*) large consumer panels.

Highly trained experts evaluate quality, and large consumer panels are used to determine consumer reaction to a product. Sensory tests performed by relatively large panels are of value in predicting consumer reactions. Evaluations by experts and trained laboratory panels are useful in quality control and product improvement studies. The trained panel can be of particular value in the assessment of product changes for which there is no adequate instrumentation. In other words, the trained panel can function as a testing instrument.

There are three fundamental types of sensory tests: preference/acceptance,

*Contribution 237, Food Research Institute, Research Branch, Agriculture Canada, Ottawa, Ontario, K1A 0C6.

[1] Sensory evaluation scientist, Food Research Institute, Research Branch, Agriculture Canada, Ottawa, Ontario, Canada K1A 0C6.

discriminatory, and descriptive. Preference/acceptance tests are affective tests based on a measure of preference or a measure from which relative preference can be determined. The panelist's personal feeling toward the product directs his response. Discriminatory tests are used to determine whether a difference exists between treatments. The panelist does not allow his personal likes and dislikes to influence his response. Laboratory difference panels can be used to determine whether there is a difference among treatments. Descriptive tests are used to determine the nature and intensity of differences.

The need for control and standardization is obvious in any type of analytical work and is especially so in sensory evaluation since it is based on psychological evaluation of physiological sensations. Factors such as testing environment, sample preparation, and method of presentation must be controlled in order to minimize their influence on judgment [*1*].[2] The selection and training of panelists is another important consideration [*2*].

Several different methods of sensory evaluation have been developed. The experimenter should be thoroughly familiar with each method, its advantages and disadvantages. The most practical and efficient method for each situation must be selected. No one method can be used universally. The experimenter must precisely determine the purpose of the test and the information he wants to acquire before selecting the method.

Triangle Test

The panelist receives three coded samples. He is told that two are the same and one is different and is then asked to identify the odd sample. This method is very useful in quality control work to ensure that samples from different production lots are the same. It is also used to determine whether ingredient substitution or some other change in manufacturing results in a detectable difference in the product. The triangle test is often used for selecting panelists.

Analysis of the results of triangle tests is based on the probability that if there is no detectable difference the odd sample will be selected by chance one third of the time.

Tables for rapid analysis of triangle test data were prepared by Roessler et al [*3*] and have been reprinted [*4,5*]. As the number of judgments increases, the percentage of correct responses required for significance decreases. Because of this, it is recommended that when only a small number of panelists are available they each perform the triangle test more than once in order to obtain more judgments.

Triangle tests were used to determine the effects of gamma radiation on the flavor of wieners and to determine the limiting dose [*6*]. Ten judges were used and each test was duplicated, giving a total of 20 judgements. With 20 judgements, 11 correct identifications are required for significance at the 5 percent level.

[2]The italic numbers in brackets refer to the list of references appended to this paper.

TABLE 1—*Triangle test results.*

Dose Mrads	Correct Judgments[a]
o.30	9
0.50	10
1.00	16[b]
0.75	12[c]
0.70	12[c]
0.65	11[c]
0.60	9
0.65	16[b]
0.60	9

[a]Total of 20 judgments.
[b]Significant $p < 0.001$.
[c]Significant $p < 0.05$.

Initial testing of wieners irradiated at 0.30, 0.50, and 1.00 megarads (Mrads) indicated that the flavor detection threshold is between 0.50 and 1.00 Mrads (Table 1). Subsequent testing of wieners irradiated at 0.75, 0.70, 0.65, and 0.60 Mrads showed that 0.65 Mrads is the lowest dose level at which a flavor change is detected. Tests were repeated at 0.65 and 0.60 Mrads and once again a difference was detected at 0.65 but not at 0.60 Mrads. These results indicate that bland wieners can be irradiated at levels of 0.60 Mrads and lower without detectable flavor differences being produced.

The results of a triangle test indicate whether or not there is a detectable difference between two treatments. Higher levels of significance do not indicate that the difference is greater but that there is less probability of saying there is a difference when in fact there is none.

Since the panelist is looking for the odd sample, the treatments should differ only in the variable being studied. All other differences should be masked. Therefore, application of the triangle test is limited to products which are homogeneous.

Paired Comparisons

A pair of coded samples is presented for comparison on the basis of some specified characteristic such as sweetness. This method has applications similar to the triangle test. Fewer samples are required and the amount of tasting is less. However, the statistical efficiency is not as great. The probability of the panelists' selecting a sample by chance is 50 percent. Roessler et al [7] published tables for the rapid analysis of the results of paired comparison tests. Two tables for paired comparisons data are presented: one to use in the case of a directional difference test when there is only one possible correct answer, called a one-tailed test; the other to use when either response can be

correct, as in preference tests, called a two-tailed test. Paired comparisons give no indication of the size of the difference between the two treatments merely whether or not there is a detectable difference.

Multiple Paired Comparisons

When there are more than two treatments to be evaluated, each must be compared with every other treatment. The number of pairs is $1/2\ n\ (n-1)$ where n = the number of treatments. This test is referred to as multiple paired comparisons.

The results of multiple paired comparisons cannot be analyzed using the same statistical tables as simple paired comparisons. At Agriculture Canada, computer programs based on the Bradley-Terry [*8*] model have been developed [*9*]. The program gives maximum likelihood estimates with necessary information for performing Tukey's test on the $\log_e$ (natural logarithm) of the estimates. The maximum likelihood estimates of the treatments always total 1.00. These values indicate how the treatments stand in relation to one another, but they give no indication of the quality of the samples.

Scheffé Paired Comparisons

Scheffé [*10*] modified the paired comparison test to ask the panelists to indicate the size of the difference detected.

This method was used to study the effect of holding temperatures after cooking on the juiciness of barbecued chickens [*11*]. The holding temperatures were 130, 140, 150, and 160°F. A panel of six judges evaluated the six pairs for juiciness. Since this method gives no indication of the juiciness of the samples, the judges were also asked to score each sample on a 6-point scale for juiciness (Table 2).

The results were analyzed by an analysis of variance. An average value (parameter) for each treatment was calculated. Once again the values of the parameters are relative; the sum of the parameters for all the treatments must total zero. The chickens held at 130°F were significantly more juicy than those held at 160°F (Table 3). As the holding temperature increased, the juiciness decreased, approximately at a constant rate.

The relation between data from paired comparison evaluation of beef tenderness and shear values on the same samples was studied using the probit model [*12*]. The extent to which shear values can be used as predictors of tenderness as assessed by a panel was determined. The number and proportion of correct discriminations associated with differences in peak force during shearing were tabulated. As the difference in maximum shear force values between two samples increased, the proportion of correct discriminations also increased. The data were examined graphically by plotting them on probability paper. The plots suggested that the probability of agreement could be described by a linear relation (probit model) $Y = A + Bx$ where A and B are

TABLE 2–*Questionnaire for Scheffé paired comparisons.*

Examine these two samples of barbecued chicken for juiciness.

1. Indicate the degree of difference in juiciness between the two samples by checking one of the following statements.

_____	is extremely more juicy than	_____
_____	is much more juicy than	_____
_____	is slightly more juicy than	_____
	no difference	
_____	is slightly more juicy than	_____
_____	is much more juicy than	_____
_____	is extremely more juicy than	_____

2. Rate the juiciness of each sample.

_____	very dry	_____	very dry
_____	moderately dry	_____	moderately dry
_____	slightly dry	_____	slightly dry
_____	slightly juicy	_____	slightly juicy
_____	moderately juicy	_____	moderately juicy
_____	very juicy	_____	very juicy

Comments:

the probit regression parameters. Using the resulting probit regression line, the size of the difference is shear force values required for panel discrimination was determined. With the type of beef used, a difference in maximum shear forces of 1 kg is associated with a 0.73 probability of discrimination; and when the difference is 2 kg, the probability is 0.90. With this type of information it is possible to make an informed decision on the size of the mean difference in instrumental readings between two treatments that is worth detecting.

Ranking

The panelist receives three or more coded samples which he is asked to rank. The results of a ranking test can be checked for significant differences by using the tables prepared by Kramer et al [*13*].

TABLE 3–*Juiciness parameter estimates for barbecued chickens at four holding temperatures.*

130°F	140°F	150°F	160°F
+0.26a[a]	+0.16 ab	-0.10 ab	-0.32 b

Standard error, 0.25

[a]Any two values not followed by the same letter are significantly different at the 5 percent level.

Prior to these tables becoming available, we transformed the ranks to scores using Table XX from Fisher and Yates. This method has been described by Larmond [*5*].

The ranking method is rapid and allows the testing of several samples at once. It is generally used for screening one or two of the best samples from a group of samples rather than to test all samples thoroughly. No indication of the size of the differences between samples is obtained. Since samples are evaluated only in relation to each, other results from one set of ranks cannot be compared with the results from another unless both contain the same samples.

Scoring

Coded samples are evaluated for the intensity of some specified characteristic. The panelist records his judgment on a graduated scale. The intervals on the scale can be labeled with numbers or with descriptive terms. The scoring method was used to evaluate the firmness and gumminess of eight varieties of spaghetti [*14*]. Analysis of variance of the results indicated that there was a significant difference among the varieties. Tukey's test was used to determine where the differences were (Table 4). From the scores, the size and the direction of the differences between treatments is evident. From the statistical analysis, the experimenter knows whether or not the differences are significant.

The descriptive terms on the scale must be carefully selected and the panelists trained so that they agree on the meaning of the terms. Objective terms (very hard) rather than preference terms (much too hard) must be used. The panelists are not typical consumers and their likes and dislikes are not considered.

There is a tendency for scales to drift in meaning with time. This

TABLE 4–*Mean firmness and gumminess scores for eight varieties of spaghetti.*

Variety	Firmness[a]	Gumminess[b]
1	4.0 c[c]	2.9 a
2	4.8 abc	3.3 a
3	4.9 abc	2.9 a
4	5.1 abc	1.9 bc
5	5.1 abc	1.6 bc
6	5.2 abc	2.0 b
7	5.9 a	1.6 bc
8	4.4 bc	1.2 c

[a] 8-point scale: 1 = extremely soft, and 8 = extremely firm.

[b] 8-point scale: 1 = no gumminess, and 8 = extremely gummy.

[c] Any two values in a column not followed by the same letter are significantly different at the 5 percent level.

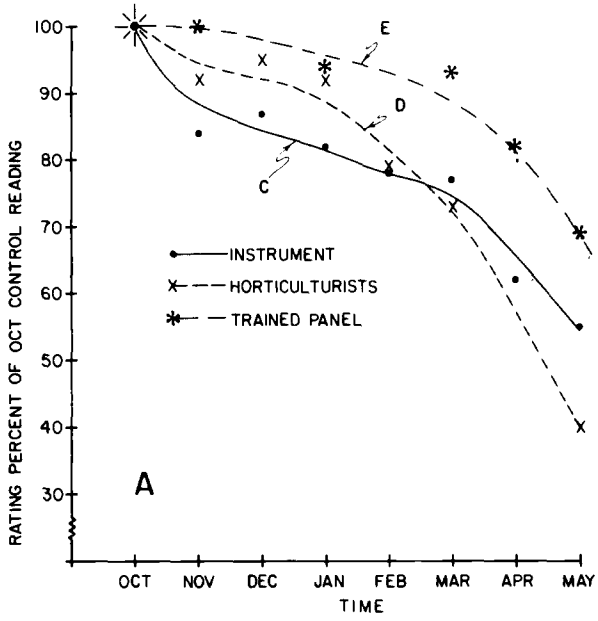

FIG. 1–*Plots of firmness values for onions against storage time as assessed by; C, instrument; D, horticulturalists; and E, trained panel.*

instability is a marked disadvantage when scoring is used in storage stability studies over an extended period.

Scoring was used to evaluate the firmness of onions stored from October to May [*15*]. Evaluations were made at monthly intervals, using an instrument, horticulturalists, and a trained panel (Fig. 1). The ratings for firmness by each method decreased with time but not at a constant rate. The instrument was calibrated before each test and so gave a constant measure whereas the sensory test did not have a reference for standardizing the evaluations. The horticulturalists who were familiar with onions tended to detect larger changes than did the trained panel, possibly because such changes were expected. The trained panel were not aware that this was a storage study, and their responses were not influenced by expectation.

Guilford [*16*] listed advantages of rating-scale methods over paired comparisons and rank order.

1. Ratings require much less time than either paired comparisons or ranking methods.
2. The procedure is far more interesting to the observers.
3. Rating scale methods have a much wider range of application.
4. They can be used with psychologically naive raters who have had a minimum of training.
5. They can be used with large numbers of stimuli.

The terms on the scale are assumed to represent equal sensory intervals,

and the scales are regarded as being linear [*4*]. These assumptions are sometimes violated. For instance, you cannot be sure that the distance between "neither like nor dislike" and "like slightly" in the hedonic scale is the same as the distance between "like very much" and "like extremely." If it is assumed that the distances between terms are equal, then the rating scale is an interval scale. Other interval scales which are well known are the centigrade and Fahrenheit scales of temperature. Distances between points on an interval scale convey information, for example, 40°F is 20 deg warmer than 20°F. However, you cannot say that 40°F is twice as warm as 20°F. Proportions cannot be obtained from interval scales.

Ratio Scaling

Another scaling method which is commonly used in physics is ratio scaling. Scales of weights and distances are examples. A distance of 40 miles is twice as long as a 20 mile distance. Ratios of the measures can be calculated. Ratio scaling is also used in sensory evaluation. The ratio measurements are usually constructed by the procedure of magnitude estimation [*17*]. The panelist is given a series of samples which vary in one characteristic (hardness). He is instructed to assign a number (say 50) to the first sample and rate each sample in relation to the first. If the second sample seems twice as hard as the first, he assigns to it the value 100; if it seems half as hard, he gives it 25. Ratio scales for more than two dozen perceptual continua have been established [*18*].

In Table 5 are results from a magnitude estimation experiment conducted by Moskowitz and Sidel [*19*]. Twenty-five panelists evaluated five variations of chocolate chip cookies using the method of magnitude estimation. The magnitude estimates were normalized by multiplying the judgements of the single panelist for each sample in each session by a constant that made his geometric mean estimate across samples equal to 1.0. The logarithms of the estimates rather than the estimates themselves were subjected to analysis after normalization.

TABLE 5–*Magnitude estimates for five variations of chocolate chip cookies.*

Sugar/Chocolate Ratio (1 = normal)	Geometric Mean	Standard Deviation
1.00/1.00	2.00	0.30
0.25/1.00	1.00	0.10
4.00/1.00	0.50	0.50
1.00/0.33	0.79	0.30
1.00/3.00	1.26	0.30

NOTE–Least significant ratio 2.0 from Moskowitz and Sidel [*19*].

The magnitude scale can measure how much more acceptable one food appears to be than another.

Descriptive Sensory Analysis

A group of highly trained panelists examine the flavor or texture of a product to provide a detailed descriptive evaluation of it. The most commonly known descriptive methods are the flavor profile [*20*] and the texture profile [*21*].

The flavor profile is the description of the flavor and aroma of a food product. The description names the perceptible factors, intensity of each, the order in which the factors are perceived, aftertaste and overall impression.

The texture profile is the description of the textural characteristics perceived in a food product, the intensity of each, and the order in which they are perceived. Mechanical characteristics are described qualitatively and quantitatively; geometrical characteristics are described qualitatively and semi-quantitatively; the type of description of fat and moisture characteristics depends on the product being studied.

Descriptive analysis is a valuable tool in difference testing and in product development work. It provides a complete description of sample differences and guides the product developer in modifying product characteristics to meet consumer demands.

The training of profile panels requires considerable time and the members must possess a high degree of motivation and interest. Once trained, however, the panel can provide thorough and reliable descriptions of products in a short time. Since the descriptive panel members work together as a group, there is the danger that forceful members will have undue influence on the results.

The descriptive sensory methods yield some quantitative data but I feel that rather than try to relate the results to instrumental data as is done with other sensory methods one should use descriptive methods to determine which characteristics of the product should be measured by instrumental methods.

Quantitative Descriptive Analysis

A method of sensory evaluation developed at the Stanford Research Institute called Quantitative Descriptive Analysis [*22*] combines descriptive analysis and ratio scaling. During preliminary sessions, the product is characterized. Samples are made up to illustrate different terms so the panel achieves agreement on which characteristics should be evaluated and the meaning of each term used. During these sessions the panel works together as a group, and discussion is encouraged. During evaluation, the panelists work individually. They record their impressions by making a vertical line across a 6-in.-horizontal line. Descriptive terms are used near the extreme ends of the scales and at the center if necessary. Following the panel the experimenter superimposes a grid dividing the line into 60 units and assigns a number between 0 and 60 to the panelist's ratings. The resultant numerical values are

then analyzed using a computer program. The unstructured line is not an interval scale, since intervals are not marked.

By using this method, the experimenter obtains quantitative data for many of the important characteristics of a product.

Conclusion

The triangle, simple paired comparisons, and ranking tests indicate whether there are significant differences. From multiple paired comparisons and Scheffé paired comparisons, the relation of the treatments to one another can be determined, but not the absolute value. The size and direction of differences among treatments is obtained from interval scaling. Ratio scaling yields proportion of differences. From descriptive analyses the experimenter obtains a qualitative description of differences.

References

[1] Larmond, E., *Food Technology*, Vol. 28, No. 11, 1973, pp. 28-32.
[2] Martin, S. L., *Food Technology*, Vol. 28, No. 11, 1973, pp. 22-26.
[3] Roessler, E. B., Warren, J., and Guymon, J. F., *Food Research*, Vol. 13, 1948, pp. 503-505.
[4] Amerine, M. A., Pangborn, R. M., and Roessler, E. B., *Principles of Sensory Evaluation of Food*, Academic Press, New York, 1965, p. 357.
[5] Larmond, E., "Methods for Sensory Evaluation of Food," Publication 1284, Agriculture Canada, Ottawa, 1970.
[6] Watanabe, W., Tape, N. W., and Larmond, E., *Canadian Institute of Food Technology Journal*, Vol. 2, No. 4, 1969, pp. 181-184.
[7] Roessler, E. B., Baker, G. A., and Amerine, M. A., *Food Research*, Vol. 21, 1956, pp. 117-121.
[8] Bradley, R. A. and Terry, M. E., *Biometrika*, Vol. 39, 1952, pp. 324-345.
[9] Dietrich, C. E., *Canadian Institute Food Technology Journal*, Vol. 27, No. 5, 1973, pp. 61-69.
[10] Scheffé, H., *Journal of the American Statistical Association*, Vol. 47, 1952, pp. 381-400.
[11] Larmond, E., Gilbert, J., and Todd, E., *Canadian Institute of Food Science and Technology Journal*, Vol. 6, No. 1, 1973, pp. 29-31.
[12] Larmond, E. and Petrasovits, A., *Canadian Institute of Food Science and Technology Journal*, Vol. 5, No. 3, 1972, pp. 138-144.
[13] Kramer, A., Kahan, G., Cooper, D., and Papavasilian, A., *Chemical Senses and Flavor*, Vol. 1, 1974, pp. 121-133.
[14] Larmond, E. and Voisey, P. W., *Canadian Institute of Food Science and Technology Journal*, Vol. 6, No. 4, 1973, pp. 209-211.
[15] Crete, R., Voisey, P. W., Bernier, R., and Larmond E., *Horticultural Science*, Vol. 9, No. 3, 1974, pp. 223-225.
[16] Guilford, P. J. in *Psychometric Methods*, McGraw-Hill, New York, 1954, p. 297.
[17] Moskowitz, H. R., Drake, B., and Akesson, C., *Journal of Texture Studies*, Vol. 3, 1972, pp. 135-145.
[18] Stevens, S. S., *American Scientist*, Vol. 48, 1960, pp. 226-253.
[19] Moskowitz, H. R. and Sidel, J. L., *Journal of Food Science*, Vol. 36, 1971, pp. 677-680.
[20] Caul, J. F., *Advances in Food Research*, Vol. 7, 1957, pp. 1-40.
[21] Brandt, M. A., Skinner, E. B., and Coleman, J. A., *Journal of Food Science*, Vol. 28, 1963, pp. 404-409.
[22] Stone, H., Sidel, J., Oliver, S., and Woolsey, A., *Food Technology*, Vol. 29, No. 11, 1974, pp. 24-34.

H. R. Moskowitz[1]

Univariate Psychophysical Functions

REFERENCE: Moskowitz, H. R., "Univariate Psychophysical Functions," *Correlating Sensory Objective Measurements–New Methods for Answering Old Problems, ASTM STP 594*, American Society for Testing and Materials, 1976, pp. 36-47.

ABSTRACT: Psychophysical scaling is discussed in terms of two basic approaches: measurement of the errors of discrimination and the erection of a scale based upon variability of judgments (Weber-Fechner approach) and measurement of magnitudes by direct subjective judgments (category and ratio scaling approaches). Each approach yields its own unique psychophysical function relating physically measured magnitudes to subjective intensities. The more direct approaches of category and ratio scaling are preferred because of their simplicity of use and because experimental findings can be translated into applications with greater ease.

KEY WORDS: sensory mechanisms, psychophysics, scale (ratio), category scaling, logarithm functions, sweetness, odor intensity, magnitude estimation

Traditional studies in psychological measurement that correlate the external, physical world with the "private world" of sensory experience rely upon one or another method of sensory scaling. Scaling is the assignment of numbers to stimuli to reflect their properties. Scaling may be as simple as giving each stimulus a number to reflect the rank order of its hardness, for example, the Mohs' scale of hardness. In an experiment using rank-order scaling, the observer is told to assign numbers to stimuli (for example, cubes of materials) to reflect the relative degree of hardness, so that if one specimen scratches another, then the harder specimen is given the higher number, etc. In more sophisticated studies the observer may be asked to assign numbers to reflect more metric relations. If Rock A seems to be 15 times harder than Rock B, and if Rock A is called (arbitrarily) 10, then Rock B might be assigned a number 15 times greater (that is, 150) to reflect this ratio of relative hardnesses.

The act of sensory scaling requires the observer to generate numbers to reflect relations among stimuli. These numbers, the scale values, have meaningful mathematical properties which are determined by the instructions given to the observer. Entire volumes have been written about the rules that underlay

[1] Head, Sensory Psychology and Methodology Unit, U.S. Army Natick Development Center, Natick, Mass. 01760; presently with MPI Sensory Testing, N.Y., N.Y. 10021.

the scaling procedures [*1*],[2] and now there are numerous schools of thought that accept different scaling assumptions. Some experimenters allow that an observer can only provide rank-order information about stimuli and that all scales derived from subjective measurement simply reflect rank-order relations among stimuli (that is, greater than versus less than). Other experimenters, especially those working in psychophysics, permit the individual to assign numbers to stimuli so that these numbers can be added, subtracted, multiplied, divided, and multiplied by a scalar, respectively. Attempts to derive equations from physically measured stimuli and subjectively scaled ones are successful only if one accepts the latter premise that the numbers that an observer generates when scaling stimuli have metric properties (addition, multiplication). Subjective-instrumental correlations are very weak if the observer is permitted to generate numbers with only rank-order values.

This paper considers two major approaches to sensory-instrumental measurement. The first is Fechner's approach, which briefly stated is that sensory measurement can best be done by measuring the observer's ability to detect small stimulus differences. The error of discrimination is relatable to a psychological unit of "intensity," and these psychological units of intensities can be correlated with physically measured magnitudes to generate a psychophysical function. The second is the more direct approach, typified by the category and ratio scales of magnitude. This approach posits that the observer himself can generate numbers to reflect sensory magnitude and that these numbers can be related to physically measured intensities to yield a psychophysical function.

Fechner–Measurement by Discrimination

In the first part of the nineteenth century the German physiologist, E. H. Weber, experimented with perceptual capacities in order to determine the minimum physical differences that man could resolve [*2*]. For example, in order to measure the discrimination capacity for weights, Weber presented the observer with two weights, 29 and 30 g, respectively. The observer was then asked to indicate which weight was heavier. The experiment was repeated, with a variety of different test weights for each 30-g standard and with a variety of standards. This procedure yielded a number that reflected the percentage increase in weight at each level of weight needed to ensure that the observer perceived the difference in weight. The same procedure was used for a variety of percepts, including weight, sweetness, brightness of lights, etc. The typical outcome of all of these experiments was that perception of stimulus differences was a relative matter. A difference of 1 g might be easily noticed when the reference weight was 6 g and the comparison weight was 7 or 5 g. The same 1-g difference became insignificant, however, when the comparison weight was 101 g and the reference weight was 100 g. This observation

[2]The italic numbers in brackets refer to the list of references appended to this paper.

suggested that the observer responded to percentage changes, not absolute changes, in sensory magnitude. In quantitative terms, this finding is expressed as the Weber fraction, $\triangle I/I$. This index of discrimination provides a percentage measure of how well small differences in stimulus magnitude are discerned.

In the latter part of that century the German physicist G. T. Fechner suggested that the sensory resolving power, which was actually a measure of the error in discrimination, might be made the basis for a unit of sensory intensity. Fechner had a conception of a scale of sensory magnitude that could be developed by summing units of discriminability. In operational terms, the experimenter would first need to obtain a measure of threshold, at which point the stimulus was just barely perceived. He would then be required to find a noticeably different stimulus and use this stimulus as the second level on the scale, with sensory magnitude of 2. The process would be repeated with the experimenter determining a stimulus level at each step just noticeably different from the one below. According to Fechner, this complete, elaborate, and tedious procedure would provide a scale of sensory intensities; any stimulus level could be assigned a number that would reflect the number of just noticeable differences (JND) that it lay above threshold.

Fechner also reasoned that the relation between sensory intensity (as measured with the JND units) and physical intensity should be logarithmic. Fechner's logic was straightforward.

1. According to Weber's finding (at least in the midrange of physical intensities), $\triangle I/I$ is constant; call this constant k.

2. When the observer makes a discrimination, Fechner surmises that a constant increase in sensory intensity has led to the discrimination. This constant increase in sensory intensity can be labelled $\triangle S$.

3. The two previous expressions can be equated to relate a psychological response ($\triangle S$) to the physical stimulus ($\triangle I/I$) that gives rise to it.

4. The difference equation $\triangle S = k' (\triangle I/I)$ can be converted to a differential equation of the form $dS = k'(dI/I)$ whose solution is the logarithmic equation

$$S = k' (\log I) + c$$

The attractiveness of Fechner's logic, coupled with the paucity of other approaches towards the relation of sensory and physical magnitudes, soon allowed Fechner's logarithmic law to hold sway. Despite the inconvenience of the experiment needed to obtain all of the parameters, the procedure does produce a scale of intensity for sensory magnitude and has been used with varying degrees of success over the past century.

At about the same time as Fechner's law of sensory magnitudes was coming into vogue, another experimenter, Plateau, suggested a minor modification of the premises and thus completely abandoned the logarithmic law. Plateau suggested that when the observer makes a discrimination he acts as a percentage measuring instrument which assesses the relative change in sensory

magnitudes, not the absolute change. Coupled with the remaining portion of Fechner's logic, Plateau deduced that a constant percentage change in physical magnitude would produce a constant sensory change in the human observer. The outcome was a power function by the following steps.

1. Discrimination conforms to Weber's law, that is, $\triangle I/I$ is constant.
2. Subjective discrimination conforms to a percentage law, which is also constant, that is, $\triangle S/S$ is constant.
3. The combination of the two preceding steps and their conversion into a differential equation leads to the equation $dI/I = k'(ds/S)$, which leads to a power function of the form $S = cI^{k''}$.

Which of these two rules is correct? Which is most tractable and useful for investigating man's perception of the world around him and which can be more useful in applications (such as setting of standards of sensory magnitude for stimulus conditions, such as room light or environmental noise)? As we will see, Plateau's suggestion ultimately has proved to be the more useful in terms of providing new approaches to sensory measurement and providing novel measures of sensory magnitudes that are useful in a variety of different situations.

Direct Scaling of Magnitude

During the early part of this century investigators interested in quantifying the perceived magnitude of stimuli by quicker and less tedious procedures than Fechner's suggested approach were introduced to the method of category scaling. A category scale is simply a series of numbers (for example, 1 to 9, 1 to 100) that reflect equal differences in sensory intensity. A value on the scale of 1 might reflect the least intense stimulus (or perhaps a stimulus that could not be detected), whereas a value of 9 or 100 at the top of the sensory scale would reflect the most intense stimulus that would be encountered. During the experiment the observer would be presented with stimuli of varying intensity, in irregular order of intensity. The observer's task would be to assign a number to each stimulus that, to him, best reflected the perceived strength of the stimulus, keeping in mind the constraint that increasingly intense stimuli were to be assigned increasingly higher numbers and that the differences in category values were to represent differences in sensory intensity. A 1-unit increase in category value was to represent the same change in sensory intensity, no matter where on the scale the 1-unit change was made. On a 9-point scale the psychological difference between a 6 and a 7 equals the difference between a 1 and a 2, etc.

The category scale values produced by the observer may be regressed against the objective instrumental measures in order to yield a psychophysical equation that relates the two measures. In linear coordinates the resulting functions that are curved, often concave downward. A logarithmic transformation of the stimulus measure usually corrects the curvature to render the

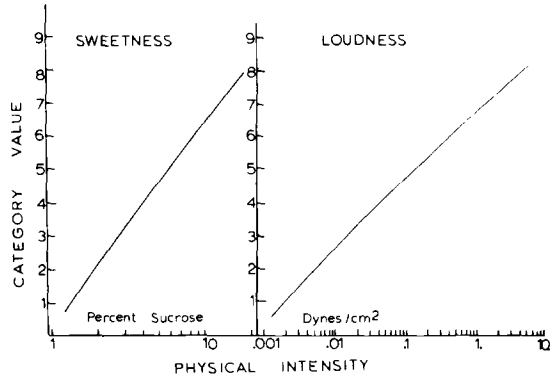

FIG. 1—*Category scale (9 Point) for the sweetness of sugar and the loudness of noise.*

function linear (or almost linear). The result is a function similar to one that Fechner had postulated

$$C = k(\log (I + I_o))$$

The value I_o is that physical intensity, a derived value, straightens out the category scale function.

Of what value are these logarithmic functions. Figure 1 shows the relation between taste intensity of sucrose and its category scale values, the relation between the amount of sound presented to an observer and the judged level of loudness. Note that with the aid of these functions the experimenter is provided with a handy instrument to assess the subjective intensity of stimuli. Note also that by asking the observer to both category estimate the stimuli and denote whether the stimulus is comfortable or uncomfortable (for example, as in the case of brightness of viewed light) the experimenter can set up "tolerance limits" of sensory magnitudes. The category scale function allows the experimenter to determine what physical intensities of the stimulus correspond to these tolerance limits.

Along a different line, the experimenter may wish to substitute the cost of the sweetening agent or the cost of the sound produced in place of the actual physical intensities. The result is a relation between sensory magnitude and cost to produce that sensory magnitude. In cases where the cost is not linearly related to the amount of sweetening agent used or the amount of sound produced, such cost-sensory intensity relations can become valuable tools to investigate economic aspects of perception.

Elegant, useful, and attractive as category scaling may seem, the scale itself lacks a number of properties and is prone to sufficient biases so that a search for its betterment is required. Some of these deficits and biases are as follows.

1. Nature does not usually measure magnitudes in terms of categories and differences, but rather in terms of ratios. We are accustomed to expressing

percentages when dealing with physical quantities. It is a far simpler task to describe one stimulus as twice as strong as another than to state that one stimulus is two units higher than another. It would be desirable to produce a scale of sensory magnitude with the same ratio or percentage properties. In that way percentage changes in the stimulus intensity can be related to percentage changes in perceived sensory intensity.

2. If the observer is provided with a scale that comprises only a limited number of categories, then he tends to bias his judgments by avoiding both the top and bottom categories, lest he should run out of available categories to describe the extremes. This is the category end effect, occurrence in scaling that Stevens and Galanter have documented extensively [*3*].

3. Quite often it appears that the observer fails to use the category scale in the intended manner. For one thing, the experimenter may provide a set of "labels" to accompany each category, which, in turn, leads the observer to use the scale as a nominal scale. The observer then selects the word that best describes the stimulus intensity, instead of using the numerical or metric properties of the scale. On the other hand, it is an act of faith to conclude that the use of the category scale is in the spirit of an equal-interval scale. Although the experimenter may set out to make the intervals, for example, between 4 and 5 equal to those between 5 and 6, etc., it is not at all clear that these intervals are used as equal intervals. In fact, the differences at the high end of the scale may substantially reflect greater differences in subjective intensity than the same numerical differences at the bottom of a scale. The subjective difference between 4 and 5 is probably smaller than that between 8 and 9, but more than that between 1 and 2. In essence, the category scale provides the experimenter with a "rubber ruler" of sensory magnitudes. This ruler seems attractive at first, but quite often the conclusions that can be reached by using it are substantially weaker because of this bias of unequal intervals of subjective magnitude.

Ratio Scaling and Power Functions

Nature measures magnitudes with a ratio scale, and much of physics relies upon the properties of proportions and percentages. Most measurements of physical quantities are expressed by numbers relative to an absolute zero (that is, similar to the Kelvin or absolute scale of temperature). Psychophysicists interested in the assessment of sensory magnitude have been enticed with the more rigorous form of measurement allowed by ratio scaling, since these numbers that emerge from the subjective measurement themselves have ratio-scale properties.

A brief history of attempts at erecting ratio scales of sensory intensity begins at Harvard University during the late 1930's and 1940's. J. G. Beebe-Center, a professor of psychology at Harvard, became interested in erecting ratio scales of taste intensity (later called the GUST scale) with the property that a taste called 50 would be subjectively twice as strong to the observer as

a taste called 25, 10 times stronger than a taste called 5, and half as strong as a taste called 100, etc. The early experiments that produced these ratio scales of subjective taste intensity were tedious and almost as time-consuming as the elaborate procedure proposed by Fechner a century before. In order to develop the scales, Beebe-Center presented the observer with a salt solution of low concentration (but still detectable) which was to represent 1 gust. The observer was instructed to sample the contents of a number of comparison glasses of sodium chloride (of varying concentration) until he found a sample that tasted twice as salty. This was denoted as 2 gusts, and the procedure was repeated until a series of salt solutions was determined with the property that these solutions lay in specific sensory ratios with respect to each other. The system was subsequently enlarged to encompass all four taste qualities (sweet, salty, sour, bitter), and sufficient experimentation and cross-comparisons of scales were done in order to ensure that 1 gust of saltiness equalled, in subjective intensity, 1 gust, for example, of sourness or bitterness, etc. An account of these experiments appears elsewhere [*4*], as do the solution concentrations for the GUST scale [*5*].

As can be done with all scales, the sensory intensities corresponding to the different stimuli can be regressed as a function of the physical intensities to yield a psychophysical function which relates each physical intensity to its expected sensory magnitude. Not only can functions be uncovered by this procedure, but physical intensities (namely, concentrations of taste chemicals) not directly scaled in the original experiment may nonetheless be assigned a "best-fitting" estimate of their intensity. The Beebe-Center GUST scale appears to conform to a power function of physical concentration, with an exponent around 1.0. Other studies of a variety of continua (for example, brightness, loudness) suggest that the power function may not have an exponent of 1.0 for all sensory modalities. Based upon results of halving and doubling the loudness of noise, the relation between sensory intensity and physical magnitude (for noise, this is dynes/cm^2) turns out to be a power function with the exponent around 0.6.

The desire for more expedient ways to assess sensory intensity by ratio scaling procedures has produced a more efficient measurement scheme, known as "magnitude estimation." In 1953 Stevens reported that when observers rated the brightness of lights and the loudness of sounds by assigning numbers to them, so that the ratios of the numbers best reflected the subjective ratios of brightness and loudness, respectively, then the results showed up as power functions. That is, the numbers assigned by the observer (S) could be related to the instrument measured intensities (I) by the function $S = kI^n$. The exponent n seemed to be unique for each sensory continuum [*6*]. For several dozen continua the exponent ranged between 0.2 (for some odorants, like benzaldehyde, diluted in a liquid medium) and 4.0 or higher (for the pain produced by electrical current).

What does this power function mean? At the simplest level, the power function from ratio scaling is nothing more than a convenient, empirical

equation that best describes the relation between a subjective percept (brightness, loudness, taste, or odor intensity) and a physical intensity that precedes it in the environment. As a descriptor the power function is parsimonious and elegant in its simplicity. Other functions, such as a polynomial function, an exponential, logarithmic or linear function might do, but the fit to the data would be oftentimes poorer. Also, the parameters of the power function for a sensory continuum (for example, loudness) often can be replicated in different laboratories, suggesting a robustness in this function.

At a slightly more complicated level, the power function is a deducible function based upon the premise that the observer transforms physical ratios of stimulus magnitude into sensory ratios and that equal stimulus ratios produce equal sensory ratios. The rule of the transformation inheres in the power function exponent. An exponent less than 1.0 means that the sensory ratio is smaller than the physical ratio. For example, if the exponent is 0.5, then a 10/1 increase in the physical intensity corresponds to a $(10/1)^{0.5}$ or 3.2/1 increase in sensory intensity. As the exponent approaches 1.0 the physical ratio and the sensory ratio more nearly equal each other. Those sensory continua governed by power functions with exponents less than 1.0 can be said to be compressive continua, large physical ratios of magnitude become compressed into smaller ratios of sensory magnitude. The range of perceived intensities for these continua is shrunk, often quite dramatically. Viscosity, hardness, and odor intensity are good examples of these compressive continua in which the physical range of variation is much larger than the sensory range of variation. On the other side lie those continua which are governed by exponents greater than 1.0. These are the expansive continua, and the story is just the opposite, for example, a continuum with exponent 2.0 produces a 100-fold increase in sensory intensity even though the instrumental measure leads us to believe that a 10-fold increment has been made. Examples of these continua are the grayness of papers, the apparent force experienced while pedaling a bicycle (versus actual expended force) and the sweetness of glucose solutions.

If the sensory intensities are plotted as a function of the physical magnitudes that precede them, then the functions appear as three separate types. Some functions curve upwards (accelerating) with physical intensity. Other functions appear linear, whereas still others curve concave downwards (decelerating). Figure 2 shows these intensity functions plotted in linear coordinates. Quite often a logarithmic transformation of both sets of coordinates (stimulus intensity, plotted on the abscissa, sensory response or magnitude estimate plotted on the ordinate) straightens out the function, as shown in right panel of Fig. 2. Those functions that accelerated, concave upward, such as the perceived shock intensity of electric stimulation now appear as straight lines with slopes greater than 1.0. In contrast, those functions that decelerated (concave downward) appear as straight lines with slopes less than 1.0, whereas those functions that were linear still appear to be straight lines, but with slopes equal to 1.0.

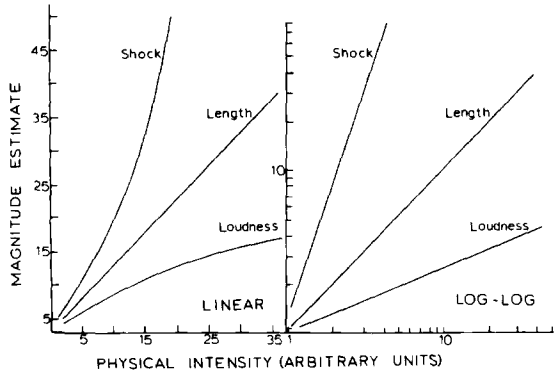

FIG. 2–*Ratio scale for three continua, plotted in linear coordinates. Power functions show up as straight lines in logarithm-logarithm coordinates.*

Although the development of such ratio scales was initiated more than two decades ago, it is only in recent years that a variety of useful outcomes have been shown to emerge from this direct form of scaling. Today the functions have become important as: (*a*) descriptions of subjective-instrumental relations; (*b*) input for models of complicated sensory processes (for example, predictions of the intensity of taste mixtures); and (*c*) predictive equations that indicate how an instrument should be recalibrated to read in terms of sensory responses (that is, a set of equations that permit the experimenter to relate instrumental measures to sensory ones and to build machines on the basis of such relations).

Some Applications of Direct Ratio Scaling

Tolerance Limits of Sensory Magnitude–Odor Intensity

Recent work on the perception of odor intensity has suggested that for a wide variety of odorants the magnitude estimates conform to a power function. In at least one case (the odor of butanol) sufficient work has been done to propose a standard method for evaluating odor intensity. Based upon studies in four separate laboratories, a function for converting butanol concentration (C) to a numerical odor intensity has been suggested [7]. This function is

$$S = 0.261C^{0.66}$$

The foregoing equation provides the experimenter with a convenient tool that converts odor intensities into numbers suitable for other purposes (for example, cross-laboratory comparisons, legislation, etc.). If other odorants are matched to butanol, then the matching level of butanol can index odor intensity for each of the odorants; and to that matching level of butanol, there may be assigned a number corresponding to its odor intensity (based

upon the butanol scale, given previously). For legislative reasons the existence of a standardized scale of odor intensity provides a convenient instrument with which to quantify the hitherto unquantified levels of environmentally noxious odors. The intensity of the odorants to the observer can be quantified, and minimum levels of odor intensity (based upon subjective judgments) can be set.

Efficacy of Different Sweetners at Varying Concentrations

Quite often manufacturers claim that their new sweeteners are several hundred or a thousand times sweeter than sucrose. These claims of intensity are usually based upon a relative measure of concentration of two sweeteners that produce the same level of perceived sweetness. For example, if 10 percent sucrose is as sweet as 0.03 percent saccharin, then saccharin is incorrectly concluded as being some 333 times sweeter than sucrose. By the method of magnitude estimation, the experimenter can obtain numbers that truly reflect the relative sweetness of saccharin and sucrose at a variety of concentrations. These functions can be plotted (for example, see Ref *8*), and the relative sweetnesses can be estimated for any two concentrations, either of sucrose, saccharin, or a pair (sucrose at level X, saccharin at level Y). The most important findings from the studies of relative sweetness versus concentration for a variety of sweet tasting substances are as follows.

1. At a fixed concentration, saccharin is almost always sweeter than sucrose. However, the concentration ranges in which saccharin and sucrose obtain commensurate sweetness ratings do not overlap; it is a meaningless question, therefore, to ask the relative "sweetness" of saccharin and sucrose at a fixed concentration. Rather, the experimenter must inquire into the relative sweetnesses of two substances, each at its appropriate concentration.

2. Saccharin and sucrose sweetnesses conform to dramatically different sweetness curves. Sucrose sweetness falls along a curve that accelerates, doubling the concentration more than doubles sweetness, whereas doubling the concentration of saccharin less than doubles sweetness. Saccharin sweetness conforms to a decelerating function. In linear coordinates the sucrose sweetness function is concave upward, whereas the saccharin sweetness function is concave downward. In logarithm-logarithm coordinates the two conform to straight lines in the middle ranges of concentration, with the sucrose exponent (for its power function) lying around 1.3 and the saccharin exponent lying around 0.6 to 0.8.

3. Other sugars, such as glucose and maltose, grow in a manner almost parallel to the growth of sucrose sweetness (except at high concentrations of sucrose where the growth in sweetness with concentration slows down quite dramatically). Most sugars conform to power functions for sweetness, with exponents around 1.3 [*9*], so that the relative sweetness of two sugars (for example, maltose, glucose) is relatively constant over an entire range of concentrations.

Through this analysis of different sweeteners the experimenters and product developments often can obtain a good idea of what levels of sucrose replace a level of glucose, fructose, etc. The experimenters need only determine the sweetness produced by the sugar to be replaced and then determine what level of the replacing sugar best produces the desired sweetness level.

Studies of Taste Mixtures and Economic Applications of Psychophysics

Rarely in the study of food tastes do we encounter simple stimuli as foods. Even sweeteners are usually used in combination, if only to minimize cost, or as in the case of artificial sweetners like saccharin to reduce any off tastes that are usually encountered. Recent studies of taste mixtures have shown that when two chemicals are mixed together that have the same taste, the mixture tastes stronger than either component alone (additivity, see Ref *10*). When the two chemicals excite qualitatively different tastes, the result is usually suppression; the taste intensity of the mixture is lower than either component. If the experimenter is able to estimate the changes in taste intensity being encountered in a mixture, then he is in the ideal position of knowing (*a*) how to compensate for the change, and (*b*) how to minimize costs when compensating for the change in taste intensities.

Applications of mixture rules in the addition of two sweeteners were suggested by Moskowitz and Wehrly based upon studies of taste mixture models [*11*]. Their approach was to consider how a pair of sweeteners would add to produce a constant total sweetness. This constant sweetness was maintained by adding together known levels of Sweetener A and Sweetener B, according to the mixture rules. However, for every level of Sweetener A, one and only one level of Sweetener B could be used in order to keep the sweetness fixed. These two levels were then transformed to costs/unit sweetener in the mixture, and the total mixture cost was computed. The dual problem was also considered in that study, maintenance of a fixed cost but optimization of sweetness subject to the fixed cost.

References

[*1*] Guilford, J. P., *Psychometric Methods*, McGraw-Hill, New York, 1936.
[*2*] Boring, E. G., *Sensation and Measurement in the History of Experimental Psychology*, New York, Appleton-Century Crofts, 1942.
[*3*] Stevens, S. S. and Galanter, E., *Journal of Experimental Psychology*, Vol. 54, 1957, pp. 377-411.
[*4*] Beebe-Center, J. G. and Waddell, D., *Journal of Psychology*, Vol. 26, 1948, pp. 517-524.
[*5*] Beebe-Center, J. G., *Journal of Psychology*, Vol. 28, 1949, pp. 411-419.
[*6*] Stevens, S. S., *Science*, Vol. 118, 1953, p. 576.
[*7*] Moskowitz, H. R., Dravnieks, A., Cain, W. S., and Turk, A., *Chemical Senses and Flavor*, Vol. 1, 1974, pp. 235-237.
[*8*] Moskowitz, H. R., *Perception and Psychophysics*, Vol. 8, 1970, pp. 40-42.

[9] Moskowitz, H. R., *American Journal of Psychology*, Vol. 84, 1971, pp. 387-405.
[10] Cameron, A. T., *Bulletin*, Vol. 9, Sugar Research Foundation, New York, 1947, p. 72.
[11] Moskowitz, H. R. and Wehrly, T., *Journal of Food Science*, Vol. 37, 1972, pp. 411-415.

Amihud Kramer

General Guidelines for Selecting Objective Tests and Multiple Regression Applications*

REFERENCE: Kramer, Amihud, "General Guidelines for Selecting Objective Tests and Multiple Regression Applications," *Correlating Sensory Objective Measurements–New Methods for Answering Old Problems, ASTM STP 594*, American Society for Testing and Materials, 1976, pp. 48–55.

ABSTRACT: The objective assessment of any sensory quality parameter must be based on its agreement with subjective "panel" evaluation. First, each sensory parameter must be accurately defined by a profile-type panel and results on a limited number of specimens correlated with several proposed objective tests. The selected procedure ($r = 0.9<$) is then correlated with difference-type panel scores on an adequate number of specimens (40+). The objective test should then be refined to improve precision and calibration. For scaling purposes, a consumer preference-type panel must be used, and the quality levels (grades) must be determined from the regression equations where panel scores are the dependent (y) variable and the objective test data are the independent (x) variable(s).

KEY WORDS: sensory mechanisms, correlation, regression analysis, subjective-objective tests

It should be emphasized at the start that the accuracy and validity of an objective assessment of any sensory quality parameter can be established only by its agreement with the direct sensory assessment of the user, not necessarily the identification of a specimen(s) in the particular sensory space. Thus in selecting an objective test, results should be highly correlated with sensory evaluation by the user. If there should be more than one type of user for a product, there may be a need for more than one objective test to predict sensory quality as it might be interpreted by the different users. For example, an objective test by which green wrap tomatoes may be evaluated for purchase by a shipper may be entirely different from those used by the retail distributor, and these in turn may differ from those used by the ultimate consumer. Evaluation of all sensory attributes is rarely required.

*Miscellaneous Publication No. 869, Contribution No. 4991 of the Maryland Experimental Station, Department of Horticulture (Food Science Program).

[1]Professor, Food Science Program, Department of Horticulture, University of Maryland, College Park, Md. 20742.

Some sensory attributes may require evaluation for one user, while others may be required for another user. In the case of the green wrap tomatoes, textural attributes may be of major importance to the shipper, appearance to the retailer, while flavor may be of interest only to the consumer [*1*].[2]

The selection of an objective method may proceed as follows [*2*]: (*a*) definition of the sensory parameter to be evaluated; (*b*) search for possible objective methods; (*c*) preliminary screening; (*d*) selection of best method(s); (*e*) improvement in simplification and precision of the selected method; (*f*) validation of the method(s); (*g*) establishing a scale; and (*h*) integration (if more than one method is required) by multiple regression analysis.

Definition of Sensory Attribute to be Evaluated

It is obvious that the user should have the primary responsibility for defining the attribute, if not directly, then by detailed and in-depth consultation with the investigator. The more precise the definition, the better the chances for the successful development of an objective test. At this point in the investigation, preferences are immaterial. What is needed here is an identification and enumeration of sharply defined parameter(s). Balanced hedonic scales combining two distinct, presumably opposite attributes should be avoided. For example, sweet-sour notes should *not* be combined into a single scale with one scored positively (+) and the other negatively (-). Rather one or more sweet or sour notes or both should be identified separately, and may be scored on a hedonic-intensity scale [*3*]. For this purpose, a well-trained, profile-type panel of users experienced in the product is of great value. The goal of such an exercise is a definition of the attribute(s) so precise that it may be possible to measure each attribute by a single instrumental or chemical procedure.

Search for Objective Methods

A good definition of the sensory attribute to be measured should of itself suggest one or more possible objective techniques for objective evaluation. A thorough survey of literature should be made not only of previous attempts to measure this and similar quality attributes for the product, but also a review of developments in related fields. Engineers, physicists, chemists, and commodity specialists should be consulted. Here is an opportunity for the researcher to apply his ingenuity, imagination, and experience which, along with the information gained from the literature and consultations, should result in a number of possible approaches. The search may well produce instruments or procedures which may be entirely satisfactory so that all that may be needed would be to compare the accuracy and precision of existing procedures and select the most reliable.

[2]The italic numbers in brackets refer to the list of references appended to this paper.

Preliminary Screening of Proposed Methods

The suitability of objective tests for measuring a sensory quality must be assessed by correlating objective test results with sensory evaluation. Since a thorough search for possible methods can produce a large number of possible tests or modifications of tests, it is appropriate to screen all these possible tests by the use of a relatively small number of specimens and a small number of panelists. The number 10 is suggested for both specimens and panelists for this preliminary screening purpose. It is helpful (although not essential, in most cases not possible) if the small number of specimens were carefully prepared to provide equally spaced sensory scores covering the entire range of sensory quality that might be encountered in the market place. If the preparation of such a set of specimens is not practical, then the specimens used should be ranked by the sensory panel rather than scored. Replicates of the same specimens may then be tested by all the proposed methods and rank correlation coefficients calculated for each. If the objective tests are performed in duplicate, it is possible at this point to screen all the methods, not only for their accuracy in predicting sensory quality but also for precision.

At this point it is usually possible to screen out most of the proposed methods because of lack of accuracy or precision or both, but most probably it will not be possible to select a best and entirely satisfactory objective procedure at this time.

Selection of Best Method(s)

Those methods that passed the preliminary screening should be tested further on a larger number of specimens so that there should remain at least 40 degrees of freedom. It is more efficient to select the specimen, not by a random survey method where a large portion of the specimens would cluster about the average quality level, but by an attempt to obtain about an equal number of specimens across the entire commercial quality range. Results of objective and sensory evaluations should be correlated as before. At this point it may be possible to select a best method if one of the tests under consideration should show very high correlation with the sensory panel that is significantly higher than a correlation between any of the other tests and the sensory panel.

It is common for such objective-subjective regressions to assume an exponential or other nonlinear relation so that it may be necessary to transform data or to fit the curve in order to obtain the best correlation coefficient. For example, the simple correlation coefficient between percent fiber in asparagus and fibrousness as measured by a sensory panel was only 0.85 as compared to a correlation coefficient of 0.94 when the percent fiber data were transformed to logarithms [*4*].

Our experience has indicated that a correlation coefficient between two

sensory panels rarely exceeds 0.94. We may, therefore, assume that 0.94 rather than 1.0 indicates practically perfect agreement between any objective and subjective test. For this reason, we assume that a correlation coefficient of 0.9 or better is an excellent indicator of the quality attribute. A coefficient of 0.8 to 0.9 is also considered satisfactory although a higher correlation coefficient is desirable. If the coefficient fails to reach 0.8, then the method is considered unsatisfactory, and that particular attribute of quality cannot be predicted with sufficient accuracy by the proposed objective test.

Another technique for removing within panel error from the objective-subjective correlation is to use two separate sensory panels and determine the correlation coefficient between those two panels. The objective-subjective correlation coefficient is then divided by the square root of the subjective-subjective coefficient. For example, a set of specimens is scored for fibrousness by two panels, and percent fiber is also determined. The mean calculated correlation coefficient between percent fiber content and fibrousness as determined by two panels is 0.9. The coefficient between the two panels is 0.85. Thus

$$\frac{0.9}{\sqrt{0.85}} = 0.977$$

and the true objective-subjective relationship, independent of intrapanel error, approaches perfection. It is possible, although highly unlikely, that such a within-panel error correction may yield a value higher than 1.0. If this should occur, the investigator should recheck the correctness of data and calculations and if found to be correct, assume a chance error similar to a rounding error. Obviously a correlation coefficient higher than 1.0 is equivalent to stating that the correlation is better than perfect.

It is possible, perhaps because of insufficient definition of the sensory attribute, that no one objective test can be found, but a combination of two or more objective tests may be used to obtain a satisfactory multiple correlation with sensory evaluation. This is frequently found to be the case when a sensory quality such as "maturity" is studied. For sweet corn, for instance, the objective tests of percent moisture, percent pericarp, and kernel size singly did not meet the requirements for a correlation coefficient of at least 0.8 with sensory evaluation when varietal and seasonal differences were disregarded. However, the multiple correlation coefficient between all three tests and sensory evaluation for maturity was found to be 0.93 [*5*].

It should be reemphasized that the use of correlation coefficients in the manner just described is valid only if the specimens used cover but do not exceed the commercial range of quality. If the specimens do not cover the entire commercial range, the correlation coefficient will probably be deceptively low (as it will be if the regression line is not fitted). If the range in quality among specimens is substantially beyond commercial limits, the

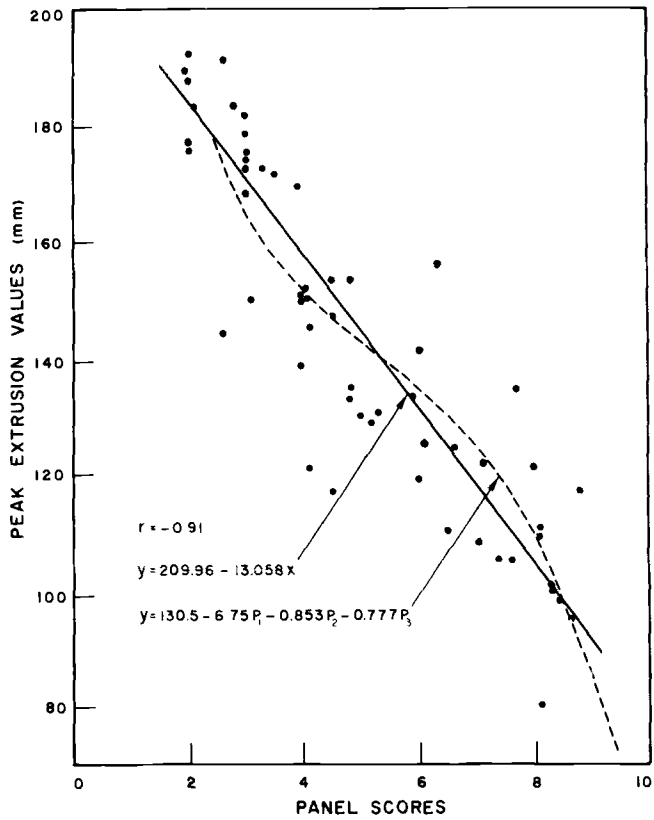

FIG. 1—*Correlation coefficient for panel scores times peak extrusion = 0.91 if all specimens are included, but only 0.68 if commercial range (panel scores 3 to 7) specimens are included.*

resultant coefficient is likely to be deceptively high [*6*]. This is demonstrated in Fig. 1 where the correlation coefficient between panel scores (subjective) for canned pea maturity and shear values was -0.91 when all specimens were included, but only -0.68 when specimens were limited to the commercial range.

Improving the Method

Assuming a satisfactory method or combination of methods are found to measure the sensory attribute, it is desirable that the selected method be as precise, simple, and rapid as possible. At this point, therefore, each step in the objective procedure should be examined carefully to see whether it can be omitted or modified with the view of improving speed and simplicity and yet retain maximum precision. Particularly in the use of instrumentation, attainment of a high level of precision is important. Thus, the first instrument manufactured can be very precise in the sense that it provides the same value

for exact duplicates. However, when a second instrument is made the results may vary from the first to the second instrument. Not only subjective but objective instrumental methods may lack in precision unless special attention is paid to the standardization and calibration of such instrumentation.

Validation

Thus far, the development of the objective test may have included special or commercial specimens whose complete history may not be known. At this point, it is necessary to establish the assurance that the method is valid not only for predicting sensory quality for those specimens that were tested but for any lots of the product, regardless of variations in ingredients or methods of manufacture, etc. It is now necessary to examine all the variables that might possibly have a differential effect on the objective method and on the subjective method. For example, it had been established on the basis of testing on a number of specimens that a measure of greenness is a good objective test for maturity of lima beans. By chance, only varieties of lima beans which fade to a near white color were used in the objective-subjective evaluations. Now, however, it is found that there are some varieties of lima beans that retain the chlorophyll pigmentation as they mature. Such specimens, therefore, would be evaluated objectively as being very young when in fact they are over mature. This step in the procedure, therefore, requires complete control over the material and a statistical design saturated with all variables that might affect the quality measurements is included. Fractional factorial designs are very useful for this purpose [*1*].

Establishing a Scale

Having validated an objective test of adequate precision and accuracy which is also rapid and simple for routine use, it is now necessary to establish a scale by which values obtained by means of the proposed objective method may be translated in terms of quality levels that will have a meaning to the user. The establishment of such a scale again requires the use of a sensory test panel, but where previous panels were of the "difference" kind, a "preference" type panel is required at this point. A hedonic scale should be used. Thus, where previously in the search and selection for an objective test the reference panel was asked to indicate, for example, the intensity of sweetness, at this point the panel should be asked whether the product is too sweet or not sweet enough. For grading purposes, it may be useful to instruct the panel to score in terms of grades, as for example "prime" for specimens having highest levels of acceptability, "choice" for the second level, "good" or "standard" for specimens which the panelist considers acceptable but not of particularly high quality, "substandard" for specimens which are usable but considered below standard. If a balanced hedonic scale is used (for individual attributes only), then it would be appropriate to use increasing positive values (+1, +2, . . .) to indicate increasingly excessive degrees sweetness; 0 would be equivalent to a

sweetness level that is considered just right; and increasing negative values (-1, -2. . . . .) would be used to indicate increasing absence of sweetness. The use of such a balanced scale would avoid considerable complications arising from those situations where very high objective values would be scored similarly to very low values. Although such relationships can be fitted mathematically, such balanced scales would simplify substantially the eventual regression equation.

In the scaling process, the selection of specimens, while important, is perhaps secondary in importance to the selection of an appropriate panel. The investigator must be certain that the panel is large enough and truly representing the buyer, not necessarily the ultimate consumer. All too frequently, for example, it is the mother rather than the infant who must be satisfied regarding baby food quality. In this case, therefore, the panel should consist of mothers, not babies. Wines are an excellent example of such esoteric performance, where the decision on sensory quality rests between the brewmaster and wine buyer with the ultimate consumer simply being told what the wine quality is. In this instance, therefore, if a wine is to be evaluated for a specific sensory quality by an objective test, for example, acidity, the objective scale for acidity should be established by reference to a panel of buyers, not consumers.

The actual scale is established from the regression equation developed from data obtained by the use of the objective test and a preference-type panel representing the buyer, who hopefully truly represents the preference of the ultimate consumer.

The actual scale is established by calculating the regress of the objective test data on the appropriate panel scores. In this instance, it is logical to use the panel scores as the dependent variable (y) and the objective data as the independent variable(s) (x).

Integration of Objective Tests

In those instances where more than one objective test is required to adequately predict a sensory attribute, similar regression analyses can be performed; but in this instance they would, of course, be multiple regressions. Here also the sensory quality values would provide the dependent (y) variable, while each objective test would provide one of the independent (x_1, $x_2 \ldots x_n$) variables. Exactly the same procedure may be followed if it is of interest to obtain an objective test or a combination of objective tests to predict in a single value a spectrum of sensory parameters.

Thus, for example, a general sensory color preference test can involve up to three objective measurements. General appearance, however, could involve many more, and general quality preference, even more. As more specific attributes of quality are included in a quality preference evaluation, however, it is usually found that certain objective tests contribute to a prediction of more than one sensory quality, so that some objective tests can be eliminated;

TABLE 1—*Summary of types of sensory panels and statistical analyses applicable at each stage of development of an objective test.*

Stage	Type	Type of Panel	Specimens	Statistical Analysis
		Panelists		
Definition	profile	5 to 10, trained	5 to 10	graphic presentation
Screening	difference	5 to 10, trained	5 to 10	rank correlation
Selection of best method	difference	10 to 20, semitrained	40+	correlation, curve, fitting
Improving the method	difference	5 to 10, trained	5 to 10	standard deviation or coefficient of variability
Validation	difference	10 to 20 semitrained	usually not less than $2^6 + 1$	analysis of variance, response surface
Scaling	preference	not trained but representative	40+	partial and multiple regression

and the overall quality preference can be satisfactorily predicted by substantially fewer objective tests than the sum of those required for the assessment of preference of each attribute individually.

Table 1 presents a summary of methods of sensory evaluation and statistical analyses to be utilized at each step in the development of an objective test.

References

[1] Kramer, A. and Twigg, B. A., *Quality Control for the Food Industry*, Vol. 2, p. 157; Vol. 1, pp. 253-283 and 357-369; Avi, Westport, Conn., 1973.
[2] Kramer, A., *Food, Drug, Cosmetic Law Journal*, Vol. 7, 1956, pp. 23-30.
[3] Mecredy, J. M., Sonneman, J. C., and Lehman, S. J., *Food Technology*, Vol. 28, Nov. 1974, pp. 36-41.
[4] Kramer, A., Haut, I. C., Scott, L. E., and Ide, L. E., *Proceedings*, American Society for Horticultural Science, Vol. 53, 1949, pp. 411-425.
[5] Kramer, A., *Proceedings*, American Society for Horticultural Science, Vol. 59, 1952, pp. 405-413.
[6] Kramer, A., *Food Technology*, Vol. 7, July 1969, pp. 66-68.

R. E. Bargmann,[1] *Louise Wu,*[1] *and J. J. Powers*[1]

Search for the Determiners of Food Quality Ratings—Description of Methodology with Application to Blueberries

REFERENCE: Bargmann, R. E., Wu, Louise, and Powers, J. J., "Search for the Determiners of Food Quality Ratings–Description of Methodology with Application to Blueberries," *Correlating Sensory Objective Measurements–New Methods for Answering Old Problems, ASTM STP 594*, American Society for Testing and Materials, 1976, pp. 56-72.

ABSTRACT: The object of this study is to relate qualitative and quantitative ratings as expressed by judges and to relate sensory responses to physical measurements and gas chromatograms. Judges were asked to rate acceptability, flavor, and mouthfeel of blueberries on a five-point scale in a darkened room. With another specimen, they were asked to judge appearance and color without, however, tasting these berries. The judges were also asked to complete a questionnaire describing, in qualitative terms, the characteristics of each specimen. Varieties were subjected to analysis by physical and chemical measurements, and gas chromatograms were prepared. Contingency analysis and analysis of variance were performed to relate qualitative and quantitative ratings. The relationship between physical measurements or chromatograms and ratings is displayed in tables of means; although these numbers can be expressed as single correlations, using a random model, a single measure based on just four products was deemed to be less informative than a tabulation of correspondence.

KEY WORDS: sensory mechanisms, analysis of variance, canonical analysis, correlations, discriminant analysis, pattern analysis, variance components, physical and chemical measurements, gas chromatograms, sensory evaluation, acceptability, flavor, mouthfeel, appearance, color, blueberries

Plan of Experiment and Analyses

Many formalistic mathematical and statistical models have been proposed to describe the relationship between physical measurements of foods, gas chromatograms, and subjective and quantified ratings by taste testers. If there is a clear match between each specimen and each replicate, statistical or even deterministic functional models can be used to attempt formal description [*1*].[2] However, where matching of individual specimens is difficult or

[1] Professor, Department of Statistics and Computer Science, research assistant, Department of Food Science, and professor, Department of Food Science, respectively, University of Georgia, Athens, Ga. 30602.

[2] The italic numbers in brackets refer to the list of references appended to this paper.

impossible, or where ratings are essentially qualitative or even categorical (nominal) in nature, such mathematical formalism may be misleading. Fortunately, in the food sciences, there has been a gradual transition from esoteric manipulative models to straightforward technique, such as regression and analyses of variance. It is informative to note the transition from earlier work (for example, Saite and Tanaka [*2*] who, much as psychologists 30 years earlier, use internal analyses, principal components, to attempt description of relationships between different sets) to exploratory statistical techniques (such as correlations between best discriminators and observable variables [*3,4*] and, finally, to regression and simple convariance models [*5*]). In the present instance, even such simplified models were deemed to be inappropriate since we dealt with perishable goods and had no way to relate specimen with specimen, let alone replicate with replicate.

In the present study, an attempt was made to keep the original experimental readings intact and apply as little mathematical formalism as necessary. Five varieties of blueberries were made available: commercial (Comm), Briteblue (BB), Southland (Sol), Tifblue (TBL), and Woodard (WRD). The amount available of TBL was insufficient to perform the physical and gas-chromatographic analyses, so that only four products were used for the study of relationships between subjective and objective measurements.

Four replications of each product were presented to 35 judges. First, in a darkened room, they were asked to evaluate flavor, mouthfeel, and acceptability of the specimen, each on a 5-point scale. Then they were asked to check, in a questionnaire (see Appendix I), qualitative descriptors of the variety that they had just tasted. Approximately 15 min after completing the first task, the judges were asked to evaluate the visual appearance and color of the specimens; the codings were different, so they had no means of relating the specimens that they had tasted to those whose visual characteristics they evaluated. Again, they were asked to check qualitative descriptors of the specimens they had rated for visual appearance.

At a later time, four replicates of the same products (but not the same specimens were subjected to gas-chromatographic analysis and the following physical determinations.

1. Drained weight, pH, and total acidity (expressed as millilitre sodium hydroxide (NaOH) consumed) were measured in the usual manner; the percent soluble solids was determined by refractometry. The viscosity of the covering syrup was measured with a viscosimeter.

2. Preliminary trials indicated that seediness would be one of the factors for determining quality. A simple flotation procedure was devised to permit separation of the seeds from the berries so that the seeds might be weighed. Five berries (3.2 ± 0.24 g) were crushed in a beaker. Then a 10 percent sodium-chloride (NaCl) solution (wt/wt) was added and the suspended seeds and pulp were alternately stirred and decanted until the skins and pulp had floated off. The ten percent NaCl solution was not high enough in specific

gravity to keep the seeds in suspension. After separation, the seeds were washed with distilled water and dried overnight at 75°C for weighing.

3. In preliminary trials, some judges commented upon the toughness of the skins; accordingly, penetrometer measurements were made on 10 blueberries from each can. An Instron instrument was used to measure the force required to puncture the skin and the total amount of work expended. The penetrometer needle had an area of 0.04 cm^2. Two peaks were evident on the time-force curve. The first peak seemed to be for compression as the berry was deformed, without penetrating the skin; ultimately, the skin would split, temporarily relieving pressure. Eventually, the probe punctured the skin, yielding a second maximum. The heights of the two maxima were recorded as centimeters and the area of the force-time curve was expressed as centimeters squared.

4. The absorbancy of the blueberry syrup, appropriately diluted, was measured at 4500 and 5500 Å in each of the following situations: (*a*) buffered at pH = 7 (coded as 7/450; 7/550; 7/max), (*b*) buffered at pH = 1 (coded as 1/450; 1/550; 1/max), and (*c*) diluted in still water (coded as H/450; H/550; H/max). In each situation, suitable dilutions were made to bring the absorbancy within the center portion of the scale. In treatments (*a*), (*b*), or (*c*), the dilution was the same for all specimens.

5. To measure reflectance, berries were pureed in a food blender and packed into a polyvinylidine chloride pouch. The pouch was then held up against a color eye. Four different filter readings were taken, X_{ce}, Y_{ce}, Z_{ce}, and X^1_{ce}. The packaged puree was then rotated 90 deg and the reflectance values were recorded again. From the color eye readings, the X_{cie}, Y_{cie}, and Z_{cie} values were calculated.

6. For the gas-chromatographic analysis, the berries were weighed out (200 g) and extracted in an extraction apparatus [*6*]. The other extract was concentrated as described by Young et al [*3*]. The stationary phase consisted of 5 percent SP 1000 adsorbed on Chromasorb AW·DMCS·HP. The mesh was 80/100. The temperature was programmed at 4.6 deg/min between 50 to 200°C. The column length was 3.66 m and the outside diameter (OD), 6.35 mm. Each peak area was expressed as the percent it was of the total area under the gas-chromatographic profile [*7*].

For the analysis of these data, several programs were employed for scaling and univariate and multivariate analyses of variance. There was a temptation to set up a variance-component model, such as let y_{ijk} represent the rating of the k'th specimen of Variety i by Judge J

$$y_{ijk} = \mu + a_i + b_j + d_{ij} + e_{ijk}$$

where a_i (variety effect), b_j (judge effect), and d_{ij} (interaction) are assumed to be random effects (Model II); then for one of the physical measurements, let z_{ik} denote the k'th replication of the measurement on Variety i. Thus

$$z_{ik} = \gamma + t_i + e'_{ik}$$

where t_i would be the random variety effect on the physical measurement. Thus, following any introductory text (for example, see Ref *8*) we could estimate "variance components" as

$$\hat{\sigma}^2 = (\text{mean square, varieties} - \text{MS interaction})/nr$$

$$\hat{\sigma}_b^2 = (\text{mean square, between} - \text{MS within})/r$$

where n is the number of judges, and r the number of replicates; and finally

$$\text{est. } cov(a_i, t_i) = \sum_{i=1}^{v} (\bar{y}_i - \bar{y})(\bar{z}_i - z)/(v - 1)$$ where v is the number of varieties.

We would then have a single measure of correlation between the rating and the physical measurements based on variety effect variances and covariances.

Now, where there are many different varieties, a good set of representatives from a relatively homogeneous class of varieties, and when safeguards are heeded (including only those judges whose intercorrelations are relatively constant), such a single number may be useful to express a relation between measurements. However, we had just four varieties on which all measurements were performed and these were certainly not representative of *all* blueberries. The temptation was great to make such a "mixed-model" analysis. We resisted this temptation and hope that whatever this analysis may lack in statistical sophistication will be made up by clarity and interpretability of results.

Selection of Judges

Since each of the five varieties was submitted to each judge four times, it was possible to perform an analysis of variance, for each judge, in order to determine whether he or she could discriminate among the five varieties on at least one of the attributes. This is a one-way analysis of variance with 4 degrees of freedom for treatments (5 varieties) and 15 degrees of freedom for error; 3.1 would be a significant F-value at the 0.05 level, and 4.9 would be significant at the 0.01 level. Thus, a decision was made to retain a judge if he attained an F-value of 3.1 or greater on acceptability or a value of 4.9 or greater on any one of the other attributes. The F-values are shown in Appendix II.

On these criteria, 23 out of the original 35 judges were retained. It is instructive to note that the strict separation between taste testing and visual appearance rating produced quite different patterns.

To find similarity of patterns between judges, a correlation matrix between judges was obtained, based upon their ratings on acceptability. By visual inspection, a subdivision into five groups was possible. Table 1 shows the

TABLE 1—*Order of preference.*[a]

Judges	Acceptability	Flavor	Mouthfeel	Appearance	Color
2, 21, 25, 35	4 2 1 5 3	2 3 4 1 5	1 4 2 5 3	2 4 3 1 5	2 1 4 5 3
6, 8, 14, 18, 31, 34	1 2 5 4 3	1 2 5 4 3	1 2 5 4 3	2 4 1 5 3	2 4 1 5 3
15, 22, 32, 33	1 2 4 5 3	1 2 5 3 4	1 2 4 5 3	2 4 1 3 5	2 1 4 3 5
4, 10, 19, 26, 29	1 4 2 5 3	1 2 4 5 3	1 4 2 5 3	2 4 1 3 5	2 4 1 3 5
5, 7, 17, 30	1 2 4 5 3	1 2 3 4 5	1 2 5 4 3	4 2 3 5 1	2 4 1 3 5
All selected	1 2 4 5 3	1 2 5 4 3	1 2 4 5 3	2 4 1 5 3	2 4 1 5 3

[a] 1 = Comm,
2 = Sol,
3 = BB,
4 = WRD, and
5 = TBL.

order in which the five varieties were rated, on five attributes, by each of the group of judges, and by all 23 judges combined. Numbers have been connected by bars if (by Duncan's multiple range test) differences were nonsignificant at the 0.05 level.

It is quite apparent from this table that the very significant variety differences are consistently due to the high superiority of Comm in the taste testing experiments and the superiority of Sol and WRD in appearance ratings.

Qualitative Descriptors and Rating Scales

In connection with the taste-testing and visual-appearance sessions, the judges were asked to check, from a list, those qualitative characteristics which best describe the varieties they had just rated. For each variety, Appendix III lists characteristics that were chosen by more than 50 percent of the raters in the taste-testing session and the visual-appearance ratings.

The connection between these descriptive characteristics and the ratings on each of the five scales was studied by means of a two-way analysis of variance. One of the factors was presence or absence of each of the descriptors, the levels of the other factor were the five varieties. For each rating scale the highest F-values (t^2 - values) for the first factor (presence-absence) have been recorded and are presented in Table 2.

It is worth noting that the following characteristics were not related to any of the attribute ratings: mashed, excessively clumped, too bland, separated from the flesh, too sweet, or too metallic. In fact, negative value judgments, such as "too thin," "too soft," "lacking," "too bland," etc., seem to be less determining for ratings than positive or neutral ones.

The face validity of the qualitative descriptors has been corroborated by

TABLE 2–*Analysis of variance, presence or absence of qualitative descriptors.*

Descriptor	F	(t^2)
ACCEPTABILITY		
Well-balanced flavor	96	
Blueberry flavor dominant	94	
Mouthfeel pleasant	72	
Firm, skin not tough	63	
Harsh, astringent	49	(neg)
Viscosity imparts pleasant mouthfeel	42	
Syrup is too thin	28	(neg)
Flavor lacking	27	(neg)
Berries too soft	23	(neg)
Seeds fine	20	
FLAVOR		
Well-balanced flavor	166	
Blueberry flavor dominant	150	
Flavor lacking	61	(neg)
Mouthfeel pleasant	59	
Viscosity imparts pleasant mouthfeel	58	
Syrup too thin	49	(neg)
Firm, skin not tough	34	
Too bland	28	(neg)
Harsh, astringent	24	(neg)
Too sour	23	(neg)
MOUTHFEEL		
Mouthfeel pleasant	95	
Firm, skin not tough	76	
Seeds fine	39	
Harsh, astringent	37	(neg)
Viscosity imparts pleasant mouthfeel	30	
Blueberry flavor dominant	26	
Well-balanced flavor	26	
Skin is tough	25	(neg)
Too many seeds	24	(neg)
Gritty seeds	24	(neg)
APPEARANCE		
Color bright and typical	41	
Appears bleached	38	(neg)
Plump and whole	31	
Lacks uniformity	18	(neg)
Color uniform	16	
Clear, attractive syrup	13	
COLOR		
Color bright and typical	75	
Color too light (but color too dark has an *F*-value of 0.7!)	28	(neg)
Color uniform	12	
Lacks uniformity	11	(neg)
Syrup clean and attractive	8	

the statistical analysis. Not a single visual appearance characteristic carried over to the taste ratings. Having the judges rate color and appearance at a session separate from that for appearance, flavor, and mouthfeel (and with the specimens coded differently) apparently avoided a halo effect or a subconscious influencing of one sensory response by that of another.

Stepwise multiple correlations between acceptability and each of the four ratings reveal, once again, the strict separation between the visual and the taste ratings. They are as follows.

acceptability versus	
all	$R^2 = 0.7009$
all except appearance	$R^2 = 0.699174$
all taste attributes	$R^2 = 0.699169$
mouthfeel only	$r^2 = 0.6232$
flavor only	$r^2 = 0.5069$
color only	$r^2 = 0.0341$
appearance only	$r^2 = 10^{-5}$

Relation to Physical and Chemical Characteristics

Experimental designs for the study of relationships between physical variables and ratings are relatively straightforward for nonperishable products. A homogeneous specimen can be subdivided, with parts of such a batch being used for the taste-testing and rating experiments and another part for the physicochemical determinations. Thus there is an exact match between each replicate given to a judge and a replicate subjected to further analyses. In perishable goods this is not possible. The only carry over from one situation to the other is the variety difference.

It has been noted in the discussion on ratings that the variety differences in the taste ratings are primarily due to the superiority of the Comm. To obtain at least a qualitative description of the relationship between subjective and objective measurements, one-way analyses of variance were performed, with the five possible responses on each attribute serving as the treatment groups and the physical characteristic of the varieties which received a rating of 1, 2, 3, 4, or 5 on one of the scales as a response variable. For those physical characteristics which produced a significant and interpretable difference between the five points on the scale, the average value for each point on the scale was calculated. In the taste variables, the average corresponding to the highest (excellent) rating has been stated in parentheses since this highest rating was given almost only to Comm and, rarely, to another one of the blueberry specimens. This last value should not be taken into consideration in the interpretation of the relationship between subjective ratings and physical attributes. Table 3 contains a record of averages of physical characteristics which produce a significant difference between ratings.

TABLE 3—*Means of significant physical characteristics, ratings.*

	1 (poor)	2	3	4	5 (excellent)	
			ACCEPTABILITY			
pH	3.14	3.12	3.17	3.25	(3.40)	
NaOH	13.4	12.8	12.5	11.1	(8.8)	
Seeds	0.18	0.15	0.13	0.11	(0.06)	
Area	1.46	1.43	1.43	1.36	(1.25)	energy to penetrate skin
Peak 02	0.96	0.97	0.96	0.83	(0.66)	pressure
7/450	0.41	0.46	0.50	0.46	(0.29)	
7/550	0.51	0.57	0.65	0.61	(0.40)	transmission
7/max	0.53	0.59	0.67	0.63	(0.40)	
1/450	0.45	0.48	0.54	0.51	(0.37)	
1/550	0.66	0.66	0.68	0.77	...	
1/max	0.77	0.81	0.92	0.90	...	
			FLAVOR			
pH	3.12	3.13	3.15	3.24	(3.30)	
NaOH	13.2	12.7	12.7	11.6	(9.8)	
Seed	0.16	0.14	0.15	0.13	(0.08)	
Peak 02	0.98	0.95	0.95	0.88	(0.75)	(pressure to penetrate skin)
			MOUTHFEEL			
pH	3.13	3.10	3.18	3.24	(3.40)	
NaOH	13.3	13.0	12.6	10.9	(9.0)	
Seeds	0.18	0.16	0.13	0.10	(0.07)	
Area	1.41	1.47	1.44	1.36	(1.23)	
			APPEARANCE			
pH	3.22	3.21	3.21	3.15	3.14	
Area	1.35	1.39	1.36	1.47	1.53	
Spectroscopy						
7/450	0.44	0.44	0.44	0.47	0.49	
7/550	0.56	0.56	0.57	0.60	0.64	
7/max	0.58	0.58	0.58	0.62	0.66	
1/450	0.48	0.47	0.48	0.51	0.54	
			COLOR			
Area	1.40	1.37	1.32	1.46	1.51	
7/450	0.44	0.44	0.43	0.47	0.49	
7/550	0.55	0.55	0.55	0.60	0.64	
7/max	0.57	0.57	0.56	0.62	0.66	
1/450	0.47	0.47	0.46	0.51	0.54	
1/max	0.83	0.82	0.81	0.87	0.93	

Table 3 shows several trivial and a few unexpected results. It is not surprising that a higher pH and lower titratable acidity (millilitre of NaOH consumed) yield better tasting varieties. Also, the fewer seeds the berry has, the higher the acceptability and the better the flavor and mouthfeel. The fact that higher absorbancy occurs with varieties rated excellent in appearance seems to indicate merely that the darker the berry, the higher its appearance

TABLE 4—*Correlations.*

Absorbancy	Seed	Area	Pk02
7/450	-0.490	0.520	0.317
7/550	-0.369	0.443	0.248
7/max	-0.412	0.481	0.289

rating. But why low absorbancy should be indicative of both poor and excellent berries in acceptability and taste, while high readings are indicative of average taste is a bit baffling. Even with the omission of the atypical fifth category, this nonmonotonic relation remains apparent. The correlation matrix between some of the absorbancy readings versus pressure of penetration and seed content may explain this relationship.

Appendix IV lists, for each variety, the mean rating and the mean values of those physical characteristics which discriminate significantly between varieties. The *F*-values are shown in juxtasposition with the mean ratings.

Appendix IV contains the same information as Table 3 for those trends which are monotonic. The three taste ratings are characterized by having the Comm high, and the two visual appearance ratings have Sol high. The Comm is higher in pH and lower in titratable acidity, seed weight, and toughness of skin; Sol is high in skin toughness and absorbancy variables. Again, an apparent relationship between the objective factors and quality of appearance is evident. Whether the relations between low values of absorbancy readings and both low and high values of taste ratings are truly related can not be determined with such a small number of different varieties.

The relation which some of the transmission readings have to taste variables can be explained, in part, by the significant correlation which one of the series (7/450, 7/550, and 7/max) exhibits versus the seed weight and the variables describing the energy required for penetration of the skin (Area and PK02) may be seen in Table 4. Also, the 1/450 reading had a correlation of -0.308 versus seed, and 0.396 versus area.

Relation to Gas Chromatograms

The problem of establishing relationships between gas chromatograms and ratings is similar to that described earlier. Again, the only link is by way of the differences among the four varieties that were subjected to this analysis. Table 5 shows the mean relative areas under gas chromatogram peaks in relation to the ratings given to the varieties. Appendix V shows the mean ratings on the five scales for each of the four varieties followed by mean relative areas of gas-liquid chromatogram peaks which showed significant variety differences. The correlation between ratings and gas-chromatogram peaks is remarkable. It is clear that peaks 03, 04, 09, 15, 17, 29 (negative), 32, 49, and 52 (negative) show convincing correlation with taste ratings, and

TABLE 5—*Means of significant relative areas under gas chromatogram peaks, specimens on which acceptability received ratings.*

	1 (poor)	2	3	4	5 (excellent)	
Pk03	0.013	0.017	0.027	0.044	(0.064)	%
Pk04	0.110	0.116	0.123	0.185	(0.280)	%
Pk08	5.1	3.7	3.0	2.7	2.7	(per 100 000)
Pk09	2.1	2.8	4.3	6.0	7.6	(per 100 000)
Pk11	0.014	0.015	0.016	0.014	(0.009)	%
Pk12	6.3	7.3	11.3	9.2	2.7	(per 100 000)
Pk15	2.2	2.4	2.9	4.0	(5.5)	(per 100 000)
Pk17	2.7	3.1	4.1	5.8	(7.7)	(per 100 000)
Pk29	0.125	0.121	0.102	0.075	(0.047)	%
Pk32	0.017	0.021	0.028	0.045	(0.068)	%
Pk39	4.5	4.7	6.4	5.6	3.1	(per 100 000)
Pk41	9.5	8.8	9.9	7.9	3.8	(per 100 000)
PK49	2.1	3.3	4.7	9.3	(15.6)	(per 100 000)
Pk52	10.5	8.1	6.3	4.2	(2.1)	(per 100 000)
Pk53	3.4	3.2	3.7	2.8	(1.1)	(per 100 000)
Pk54	1.9	2.1	2.7	2.1	0.8	(per 100 000)

NOTE—Mean values for flavor and mouthfeel are almost identical.

11, 12, 39, 41, 53, 54 have a close relationship to appearance ratings. This relationship is brought out even more strongly in Table 6 where the mean peak area for each of the five scale points of the rating scales has been tabulated.

As in the comparison with physical characteristics, the mean values for acceptability in the highest (excellent) category should not be interpreted. The great dominance of the Comm in this category tends to confound this highest rating scale point (enclosed in parentheses). It must also be noted that

TABLE 6—*Appearance.*

	1 (poor)	2	3	4	5 (excellent)	
Pk03	0.023	0.025	0.030	0.038	0.037	%
Pk04	0.170	0.169	0.165	0.110	0.093	%
Pk08	3.6	4.2	3.7	1.8	2.2	(per 100 000)
Pk09	3.1	3.3	4.1	6.1	6.0	(per 100 000)
Pk11	0.011	0.011	0.013	0.020	0.021	%
Pk12	1.5	2.7	5.8	18.6	20.7	(per 100 000)
Pk15	3.0	3.1	3.3	3.3	3.1	(per 100 000)
Pk17	3.7	3.9	4.4	5.1	5.0	(per 100 000)
Pk29	0.115	0.108	0.099	0.082	0.081	%
Pk32	0.032	0.032	0.035	0.032	0.029	%
Pk39	2.4	3.0	4.2	9.3	10.2	(per 100 000)
Pk41	5.5	6.5	7.5	12.3	13.8	(per 100 000)
Pk49	6.5	6.4	6.8	5.4	4.2	(per 100 000)
Pk52	7.0	7.8	6.8	4.1	4.9	(per 100 000)
Pk53	1.8	2.1	2.6	4.8	5.4	(per 100 000)
Pk54	1.1	1.2	1.7	3.8	4.1	(per 100 000)

NOTE—Mean values for color are very similar.

F-values for differences of means between rating-scale points are not valid. Appendix V shows the only valid *F*-values for differences.

The following areas under the peaks show significant correlation with acceptability (taste) and appearance (visual): 4, 8 (decreasing), 9, 17, 29 (decreasing); 49 (strongly increasing with taste, slightly decreasing with appearance); and 52 (decreasing).

The following areas correlate with acceptability (taste) but not with appearance: 3, 15, and 32.

The following areas show strong relation with visual appearance of the varieties, but no association with taste ratings: 11, 12, 39, 41, 53, and 54.

Summary

It has been demonstrated that in the assessment of the quality of blueberries, verbal expressions checked by judges and responses on five rating scales (three for taste and two for appearance) are so well related to each other that they can be used interchangeably. Since a specimen to specimen matching between taste assessment and physicochemical determinations is next to impossible with perishable items, an attempt has been made to base a study of relationships on the variety differences only. Even with this limited information it has been possible to establish a significant relationship between low acidity and low seed weight on the one hand and acceptability by taste testing on the other. Also, blueberries with firmer or tougher skin and higher absorbancy readings were judged superior in appearance. The relationship between 16 gas-liquid chromatogram peak areas and taste is so convincing that it would seem that reliable predictions of taste quality could be based upon a study of gas-chromatograms alone. Gas-liquid chromatogram peak areas and appearance also were highly correlated. The correlations between taste ratings and visual appearance are zero; there is no intrinsic connection between them. It seems important to conduct such experiments in darkened rooms for the taste testing experiments and independently of the visual appearance ratings.

APPENDIX I

Judges' Questionnaires

Date _____ Name_____

Examine the five plates of blueberries and check off the word description which most closely applies.

Acceptability

Excellent (would be willing to eat gladly)	____	____	____	____	____
Good (find quite desirable)	____	____	____	____	____
Fair (acceptable)	____	____	____	____	____
Poor (would finish a serving)	____	____	____	____	____
Undesirable (would not finish a serving)	____	____	____	____	____

Flavor

Delightful, balanced, full blueberry flavor	____	____	____	____	____
Pleasant blueberry flavor, only a trace of blandness, harshness, astringency, or other imbalance[a]	____	____	____	____	____
Good blueberry flavor[a]	____	____	____	____	____
Fairly good blueberry flavor[a]	____	____	____	____	____
Poor blueberry flavor[a]	____	____	____	____	____

Mouthfeel

Berries plump, neither too tough nor too soft, skins not too tough, not excessive number of seeds or tough seeds, generally pleasant mouthfeel	____	____	____	____	____
Mouthfeel quite pleasant, berries a trace soft or tough, slight seediness or tough seeds, slightly tough skins or other flaw[a]	____	____	____	____	____
Mouthfeel good[a]	____	____	____	____	____
Mouthfeel fair[a]	____	____	____	____	____
Mouthfeel poor[a]	____	____	____	____	____

Having rated the five specimens for acceptability, flavor, and mouthfeel, go into the taste tasting room and rate the berries for the two qualities below.

Appearance

Highly uniform in appearance, in size, no or few splits, acceptable liquid/solid ratio, little clumping of berries	____	____	____	____	____
Quite uniform in appearance[a]	____	____	____	____	____
Noticeable variation in appearance[a]	____	____	____	____	____
Definite variation in appearance[a]	____	____	____	____	____
Objectionable variation in appearance[a]	____	____	____	____	____

Color

Highly desirable uniform color	____	____	____	____	____
Good color and quite uniform[a]	____	____	____	____	____
Fairly good color and uniform[a]	____	____	____	____	____
Undesirable off shade or variation[a]	____	____	____	____	____
Definite off shade or variation[a]	____	____	____	____	____

[a]Except for the first category, specify the flaw(s), for example, for appearance: berries split, too much liquid, clumping of berries, etc; for color: off shade, excessive variation among individual berries, color not typical of canned blueberries.

Name_____ Specimen Color Code_____ Date_____

Please check the most appropriate response below. You may have more than one response to the statement. Do one sheet for each color coded specimen.

The berries look whole and plump _____, mashed _____, shriveled _____, excessively clumped _____, too many skins are separated from the flesh.

The syrup is clean and attractive looking _____, cloudy _____, contains excessive debris (particles of skin, seed particles) _____.

The color is bright and typical of blueberries _____, it is uniform _____, the color of the berries lacks uniformity _____, the color is too dark _____, bleached or too light _____.

The viscosity of the syrup imparts a pleasant mouthfeel _____, the syrup is too thin _____, too thick _____.

The berries are firm and the skin is not too tough _____, the berries are too soft _____, too hard _____, the skin is tough _____.

The products contain too many seeds _____, the seeds or seed particles are tough or gritty _____, there are not an undue number of seeds and they are not too hard _____.

The feel of the product in the mouth is pleasant _____, too pulpy _____, too unctuous (salvy) _____, harsh, astringent, or puckery _____.

The blueberry flavor is dominant with the sweetness complementary to it _____, the product is too sweet _____, too sour _____, too bitter _____, too metallic _____, too bland _____.

The flavor is well balanced and redolent of blueberries _____, the flavor is weak _____, lacking _____, atypical _____.

APPENDIX II

F Values for Product Differences for 23 Selected Judges

Judge	Acceptability	Flavor	Mouth Feel	Appearance	Color
7	18.3	8.4	9.1	2.6	3.8
14	16.1	11.2	25.1	10.9	15.9
5	14.8	8.8	8.8	1.3	9.1
17	14.1	3.8	7.5	3.1	11.5
30	12.1	11.2	5.6	2.5	1.7
33	9.3	2.6	10.8	0.9	1.3
25	8.0	7.6	8.4	2.4	2.0
32	7.4	7.4	5.7	4.7	5.1
4	7.3	5.1	9.5	3.3	10.5
8	5.6	10.5	8.3	0.9	2.1
34	5.0	3.8	4.5	5.7	5.0
29	4.2	3.6	10.0	7.3	3.9
10	3.9	2.3	5.6	16.1	7.8
22	3.8	6.2	3.0	3.5	21.0
15	3.8	3.6	3.0	2.5	13.1
31	3.7	0.3	1.8	1.6	0.8
26	3.5	1.5	3.3	7.9	21.3
6	3.4	3.4	5.2	2.2	3.9
19	2.7	0.6	8.2	3.3	4.0
21	1.7	1.2	3.3	5.7	16.0
18	0.7	1.7	0.4	4.2	6.8
35	0.4	1.1	0.4	3.7	7.4
2	0.4	0.4	2.6	4.9	1.9

APPENDIX III

Variety	Descriptors	%
TASTE		
Comm	syrup imparts pleasant mouthfeel	83.7
	firm and skin is not too tough	71.2
	seeds not too numerous, not too hard	94.6
	blueberry flavor is dominant	52.2
	flavor is well balanced	71.7
Sol	syrup imparts pleasant mouthfeel	66.3
	skin is tough	67.4
BB	syrup imparts pleasant mouthfeel	59.8
	skin is tough	71.7
	contains too many seeds	61.9
	seeds are tough or gritty	69.6
	pulpy	54.3
WRD	syrup imparts pleasant mouthfeel	63.0
	seeds are tough or gritty	54.3
TBL	syrup is too thin	52.2
	skin is tough	52.2
APPEARANCE		
Comm	mashed	64.1
	shriveled	54.3
	syrup is clean and attractive	63.0
Sol	shriveled	52.2
	color is bright and typical	61.9
	syrup is clean and attractive	76.1
BB	shriveled	71.9
	syrup is clean and attractive	72.8
	lacks uniformity	70.6
WRD	syrup is clean and attractive	73.9
TBL	shriveled	65.2
	syrup is clean and attractive	69.6
	lacks uniformity	53.3
	bleached or too light	58.7

APPENDIX IV

Means of Products

Variety	Acceptability	Flavor	Mouth Feel	Appearance	Color	pH	NaOH	Seed	Area
Comm	4.42	4.33	4.56	2.65	3.67	3.48	8.0	0.058	1.17
Sol	2.98	3.43	2.85	3.79	4.17	3.18	12.4	0.093	1.81
BB	2.06	2.88	2.42	2.67	2.58	3.10	14.6	0.279	1.58
TBL	2.33	2.67	2.02	2.52	2.46	3.13	12.4	0.107	1.33
F-values	...	...	...	...	...	3.7	13.8	40.5	3.1

	Pk02	7/450	7/550	7/max	1/450	1/550	1/max
Comm	0.60	0.23	0.33	0.33	0.31	0.52	0.59
Sol	1.23	0.86	1.15	1.20	0.96	1.29	1.62
BB	0.93	0.28	0.31	0.33	0.32	0.49	0.55
TBL	0.88	0.36	0.46	0.46	0.38	0.56	0.66
F-values	4.23	27.7	23.8	25.4	20.5	14.9	21.6

	H/450	H/550	H/max	H/Ratio
Comm	0.31	0.52	0.52	0.61
Sol	0.86	1.10	1.18	0.80
BB	0.46	0.61	0.64	0.75
TBL	0.44	0.58	0.61	0.77
F-values	16.7	7.2	9.2	6.7

APPENDIX V

Mean Relative Areas under Peaks of Gas Chromatograms

Variety	Acceptability	Flavor	Mouth Feel	Appearance	Color	Peaks 03	04	08
Comm	4.42	4.33	4.56	2.65	3.67	0.071	0.313	3.1[a]
Sol	2.98	3.42	2.85	3.79	4.17	0.042	0.065	0.9[a]
BB	2.06	2.88	2.42	2.67	2.58	0.008	0.117	9.5[a]
TBL	2.33	2.67	2.02	2.52	2.46	0	0.110	0
F-values	...	...	...	...	...	5.12	3.18	3.52

Variety	Pk09	Pk11	Pk12	Pk15	Pk17	Pk29	Pk32
Comm	8.1[a]	0.007	0.63[a]	6.0[a]	8.3[a]	0.035	0.075
Sol	7.4[a]	0.025	28.5[a]	3.3[a]	5.6[a]	0.069	0.029
BB	0.5[a]	0.009	1.3[a]	1.9[a]	2.0[a]	0.129	0.012
TBL	1.6[a]	0.015	2.2[a]	1.8[a]	1.8[a]	0.158	0.016
F-values	4.62	5.63	3.95	7.61	6.08	11.6	11.5

Variety	Pk39	Pk41	Pk49	Pk52	Pk53	Pk54
Comm.	2.3[a]	2.8[a]	17.6[a]	1.9[a]	0.55[a]	0.38[a]
Sol	13.3[a]	16.6[a]	3.8[a]	3.0[a]	6.7[a]	5.4[a]
BB	2.8[a]	10.7[a]	0.35[a]	3.7[a]	3.5[a]	0.88[a]
TBL	2.2[a]	3.8[a]	3.2[a]	3.7[a]	1.4[a]	1.7[a]
F-values	6.87	4.64	5.74	7.76	5.44	7.26

NOTE–Values in percent of total area.

[a]Values in per 100 000.

References

[1] Persson, T., von Sydow, E., and Akesson, C., *Journal of Food Science*, Vol. 38, 1973, pp. 682-687.

[2] Saite, N. and Tanaka, T., *Journal of Fermentation Technology*, Vol. 45, No. 11, 1967, pp. 1023-1033.

[3] Young, L. L. Bargmann, R. E., and Powers, J. J., *Journal of Food Science*, Vol. 35, No. 3, 1970, pp. 219-223.

[4] Dravnieks, A., Reilich, H. G., Whitfield, J., and Watson, C. A., *Journal of Food Science*, Vol. 38, 1973, pp. 34-39.

[5] Kosaric, N., Duong, T. B., and Surcek, W. Y., *Journal of Food Science*, Vol. 38, 1973, pp. 369-373.

[6] Milutinovic, Leposava, Bargmann, R. E., Chang, Kun-Yu, Chastain, Margaret, and Powers, J. J., *Journal of Food Science*, Vol. 35, 1970, pp. 224-228.

[7] Powers, J. J. and Keith, E. S., *Journal of Food Science*, Vol. 33, 1968, pp. 207-213.

[8] Peng, K. C., *The Design and Analysis of Scientific Experiments*, Addison-Wesley, 1967.

M. E. Wells[1]

Response Surface Methodology and Subjective Data

REFERENCE: Wells, M. E., "**Response Surface Methodology and Subjective Data,**" *Correlating Sensory Objective Measurements–New Methods for Answering Old Problems, ASTM STP 594*, American Society for Testing and Materials, 1976, pp. 73–80.

ABSTRACT: The use of quadratic models in two or more variables to describe a subjective response is discussed. Attention is given to the principles of experimental design as well as to any special problems incurred by the use of ranking or rating data as a response. An example taken from the textile area serves as an illustration.

KEY WORDS: sensory mechanisms, factor, response, experimental design, quadratic model, correlation, rating scale, ranking, paired comparisons

Experiments are ordinarily conducted to define the effects of variables on some measured output, or response. If the variables are quantitative, then the response y can be described mathematically as some function of them: $y = f(x_1, x_2, \ldots, x_n)$. The exact form of this function is not often known, but the data collected in the experiment can be used to formulate an empirical model which will serve as an approximation to the exact model and which will adequately describe the response. Polynomial models have been found to be very effective approximations in many cases.

When a polynomial model is to be used to describe a subjective response, such as fabric luster, or preference for a product, special problems arise. The luster of a fabric is not a numerical quantity, and it must be made such before any equation can be formulated.

This paper comprises three parts: (*a*) the planning stages; the designing of an experiment which will produce the desired results; (*b*) methods which can be used to provide a numerical measurement of the subjective response; and (*c*) an example to illustrate the principles and techniques discussed, as well as the presentation of results.

[1] Statistician, Engineering Department, E. I. du Pont de Nemours and Company, Wilmington, Del. 19898.

Planning Stages

The general plan is to collect sufficient data to develop an equation. This equation will express the measured quantity or response as a function of certain variables whose values can be controlled by the experimenter. The measured quantities, or the y's of the equation, are called dependent variables. They may be "objective" quantities such as process yield, or yarn strength. Or they may be "subjective" quantities: fabric luster, softness, consumer preference for an item, or how well an experimental yarn weaves. The controllable quantities, the x's of the equation, are called independent variables. They may be such things as temperature, yarn twist, or fabric blend level. Each independent variable can be set to specific values by the experimenter and changed at will to determine whether or not it produces a change in the response. The formulation of an equation does not just happen; careful planning is required to ensure that the data will allow it. For instance, in order for the data to support a quadratic model, experiments must be run at three values, or levels, of each independent variable: a low level, a high level, and a midpoint. The reason for this is that the polynomial approximation will contain a squared term in each x variable. If the response is measured at only the extreme values of x, the best that the experimenter can do is to draw a straight line between the two measured values to describe how the response changes with x. If the response is measured at a third value of x, the experimenter has information which will tell him whether or not the response should be described by a curve and how that curve should look. Note that if the third value is placed too close to either extreme, it becomes useless. See Fig. 1.

The actual values of the high and low levels must be chosen with care. The measured values of the response will not be exact; they will be contaminated with whatever error is contained in the measuring process. In the case of subjective data, this error could be due to person-to-person variability, subject fatigue, or ambiguous definition of the response itself. If the extreme values of x are too close together, a change in the response produced by shifting from low x to high x would be very small. The detection of this change, if possible, would require large numbers of indefatigable subjects. The values of x should be set far enough apart so that the experiment will produce items variable enough for the subject to rank or rate.

Methods

Which experimental runs should be made in order to find out how the variables are affecting the response? The most obvious answer is to make one run for every possible combination of variables and levels. This requires nine runs when there are two independent variables, 27 when there are three, and 3^n runs when there are n variables. Just as it is possible to err by not collecting enough information, it is possible to err by collecting too much. The results will certainly be valid, but time and money will be unnecessarily

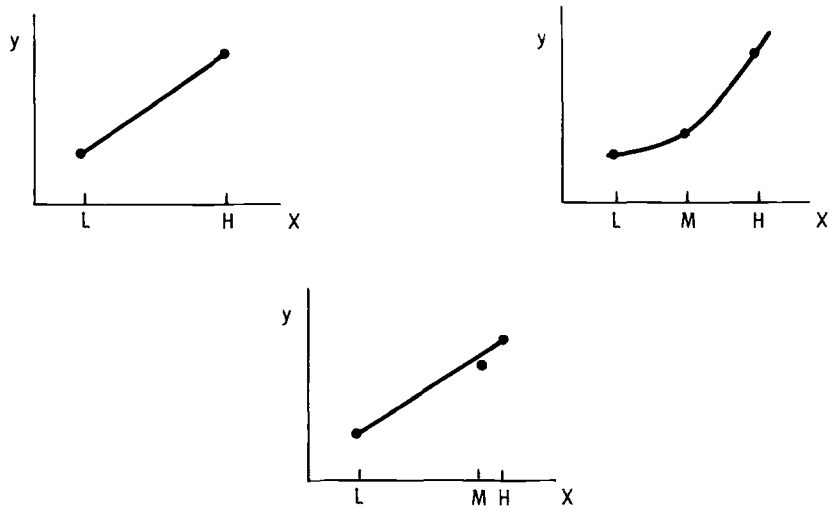

FIG. 1—*(a) Factor at two levels; (b) factor at three levels, evenly spaced; and (c) factor at three levels, unevenly spaced.*

spent in running every possible combination of variables and levels as the number of variables increases. In most cases a manageable number of runs (a subset of the full set) can be made and the equation developed from the data used to predict what the others might have been. Just any subset of runs will not suffice, however. They must be chosen so that the changes made in any one variable are not associated with the changes in any other. This ensures that the effect on the response computed for each variable reflects the changes in that variable alone.

The experimenter must also take care that the runs are not all localized in one section of the experimental region. If the equation developed from the data is to describe the behavior of the response over a certain region, then data must be collected in all parts of that region.

There are a number of experimental plans which meet these and other criteria for good experimental design. One, called a Box-Behnken design in three variables, requires 15 runs instead of the full 27. It is the design used in the example to be discussed later.

As has been noted previously, the responses, or dependent variables, must be numerical quantities for the formulation of an equation. When the responses are physical measurements or instrument readings, there is no problem. Subjective responses such as product appearance or fabric uniformity cannot be measured with a meter stick or read from a dial. There is, however, a variety of ways in which numerical values can be assigned to subjective responses, enabling them to be treated in the same fashion as objective data.

One way to arrive at a set of numbers describing a subjective response is to use a rating scale. Item one may be assigned an arbitrary number by the

subject, who will then assign all other items numbers which describe them relative to item one. This method is often not effective, since it depends heavily upon the subject's recollection of the first item. His perception of that item may change appreciably as he examines the others. To make the subject's task easier, standards may be used to construct a scale ranging from good to bad, soft to harsh, etc. The subject is asked to place each item at some point along the scale. Using standards has the advantage of making the scale uniform for all subjects. The more standard items used, the less interpolation required of the subject.

Rating scale data has the advantage of being able to be used directly, as it is collected from the subject, without further analysis. In addition to providing the necessary numerical quantities, it also estimates some degree of difference between items. It does, however, require two things: first, thoughtful construction of the scale, rather than a quick, arbitrary assignment of standards by the experimenter, and second, for best results, the subjects should be trained in using the scale.

Ranking data can also be used for the same purpose. The end product is a complete ranking of all the items resulting from the experimental runs from, for example, most to least lustrous. There are a number of ways to obtain the ranking. The most straightforward is to rank the entire sample on the basis of the property in question. This is quite easy to do when there are five or six items, but becomes progressively more difficult as the number of items increases. Like rating scales, this method of ranking produces data which can be used directly. The experimenter need only assign consecutive numbers to the items in the order in which they are ranked. It does not, however, allow for the fact that adjacently ranked items may not all be equal distances apart; some pairs of items may be closer together than others.

For medium sized groups the method of paired comparisons can be used. The subject is presented with pairs of items and, for each pair, is asked to note which he prefers or which is more lustrous. Since the number of possible pairs multiplies quickly as the number of items increases, it is often a good idea to present the items in triples. By answering two questions (which is most lustrous? which least?), the subject actually makes three paired comparisons per triple.

The example in Table 1 below is a six-item design requiring ten triples. It has an extra feature; every pair appears twice, thus providing a check on the consistency of the subject. Note that every possible pair of items appears in the design. If pairs of items are to be duplicated, each pair should appear the same number of times. This type of ranking data cannot be used directly, and its conversion to a usable form is easiest when the design is balanced.

Data for any subject can then be summarized in a matrix such as the one shown in Table 2. In this test design, the subject saw every pair of items four times. The numbers in the matrix tell how many times each row item was more lustrous than each column item. The row sums are the numerical values used in formulating a model. If a subject is perfectly consistent, the row sums

TABLE 1–*Design for ranking six items.*

Triple	Items		
1	1	2	5
2	2	6	1
3	1	4	3
4	6	1	3
5	1	5	4
6	3	2	4
7	2	5	3
8	4	2	6
9	6	5	3
10	4	6	5

are evenly spaced, as shown in Table 2, and the paired comparison method, like other ranking methods, does not show degrees of difference between items. When the subject is not perfectly consistent, the situation changes. See Table 2. Here, one item stands out from the others. If a subject is unable to tell differences between certain items but is forced by the paired comparison method to make a choice, his confusion will be evident in the row sums, which will be close together for similar items.

The paired comparison method can be used to rank any number of items, including the fairly large groups which can result from response surface designs. But the number of pairs multiplies quickly as the number of items increases. This system becomes very tedious and requires a large number of samples. For larger groups, ranking can be done by dividing the set into halves (more and less lustrous) and into halves again and again until the subgroups become small enough to rank. Another method is to have the subject choose the best and the worst from the set of items, put them aside, choose the best and worst from the remaining items, and continue in this fashion until all items have been chosen. When these methods are used, it is important that the subject be asked to check his ranking and change the position of any items that seem to be out of order when he views his work as a whole.

TABLE 2–*Paired comparison data summary.*

Item	1	2	3	4	Sum
		Consistent Subject			
1	0	4	4	4	12
2	0	0	0	0	0
3	0	4	0	4	8
4	0	4	0	0	4
		Inconsistent Subject			
1	0	4	4	4	12
2	0	0	2	2	4
3	0	2	0	2	4
4	0	2	2	0	4

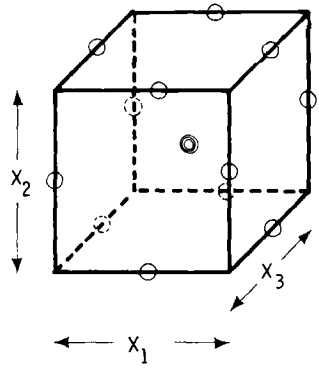

FIG. 2—*Box-Behnken experimental design, three variables at three levels.*

Ranking methods in general have the advantages of requiring no recollection of what went before and no reference to standards. They usually give better data than the rating scale methods when inexperienced subjects are used. They usually do not show degrees of difference between items. The paired comparisons method does not confuse the subject by asking him to look at more than two or three items at one time, but it does require additional analysis to convert the data to usable form.

Illustration of Principles and Techniques

The problem which will illustrate some of the principles discussed above is a three-variable study in the area of fabric aesthetics. Two resins were applied to polyester-cotton fabrics of the type used in children's sleepwear. It is always important that a fabric be able to meet certain performance standards and at the same time retain the aesthetics which it needs for its particular end use. In the case of children's sleepwear, softness is a primary concern. The aim of the experiment was to find a combination of the independent variables which is optimum both in terms of meeting performance standards and keeping the fabric soft. Only the determination of the effects of the variables on softness will concern us here.

The independent variables in the experiment were three: the amount of polyester in the polyester/cotton blend and the two resins applied to the fabric. The experimental design used was the Box-Behnken design seen in Fig. 2. Note that each variable appears at three levels to allow estimation of the squared terms and that the design requires only 15 runs. Of these runs, 13 are spread throughout the experimental region; two are repeats of the run located at (0,0,0). Before the experiment could be run, the design was translated from the zeros and ones which designate the levels of the variables to the actual values of the variables. The amount of polyester in the blend ranged from a low of 25 percent to a high of 75 percent, with a midpoint at 50. The other two variables, called Resin A and Resin B, ranged from 12 to 20 percent, with

TABLE 3—*Fabric softness data.*

Run	X_1	X_2	X_3	Y
1	50	16	16	2.25
2	50	20	20	1.00
3	75	20	16	1.92
4	50	20	12	3.00
5	50	12	20	1.75
6	75	12	16	2.25
7	25	20	16	3.42
8	50	16	16	1.83
9	50	12	12	3.67
10	25	12	16	3.92
11	75	16	12	2.00
12	75	16	20	1.25
13	25	16	20	2.33
14	25	16	12	3.42
15	50	16	16	2.00

16 as the midpoint in both cases. The order in which the experimental runs were made was deliberately random so that any systematic error, due to the order of the trials, would not affect the data. The center points were run at the beginning, the middle, and the end of the experiment.

The 15 treated fabrics resulting from the experimental runs were to be ranked on the basis of softness. Although it was quite clear that they represented varying degrees of softness, it was also clear that many of them were too much alike to allow a distinct one-to-fifteen ranking. The problem was solved by asking the subjects to place the fabrics into four groups as they ranged from harsh to soft. The numbers one (harsh) to four (soft) were assigned to the groups, and each fabric was assigned a number according to its group. The resulting data are averages of the work of twelve subjects, all of whom are experienced in handling fabrics. See Table 3.

These data were used to fit an equation expressing softness as a function of the amount of polyester, Resin A, and Resin B. Tests of significance showed that the amounts of polyester and one of the resins do have a significant effect on the softness of the fabric. Accordingly, the plot in Fig. 3 was made from the equation which describes the response over the full range of these variables. The amount of the other resin, which has no effect on softness, is held constant at its midpoint. It is easily seen that the softest fabric is found in the lower left-hand corner of the plot, at the lowest levels of polyester and Resin B. Had Resin A also produced a significant change in the response, three plots (one each at the high, mid, and low values of Resin A) would be necessary to describe the results.

The design and execution of response surface experiments require thoughtful planning no matter whether the data collected are quantitative or qualitative. The measurement of the subjective response requires special techniques, some of which have been reviewed here. The developing of a response surface

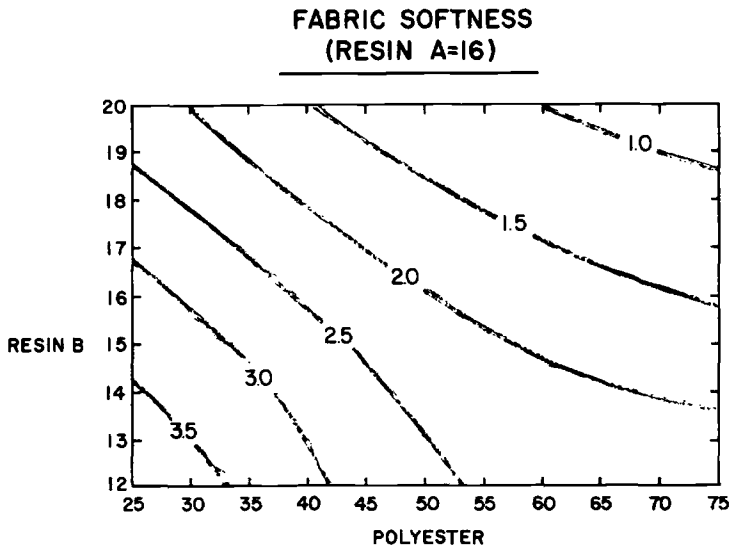

FIG. 3–*Softness as a function of amount of polyester and amount of Resin B.*

model from the data permits a graphical display of the results, which in turn provides good guidance in selecting variable settings to produce a desired response.

Bibliography

Cochran, W. G. and Cox, G. M., *Experimental Designs*, Wiley, New York, 1957.
Draper, N. R. and Smith, H., *Applied Regression Analysis*, Wiley, New York, 1966.
Goulden, C. H., *Methods of Statistical Analysis*, Wiley, New York, 1952.

G. E. Stungis[1]

Overview of Applied Multivariate Analysis

REFERENCE: Stungis, G. E., **"Overview of Applied Multivariate Analysis,"** *Correlating Sensory Objective Measurements–New Methods for Answering Old Problems, ASTM STP 594*, American Society for Testing and Materials, 1976, pp. 81–96.

ABSTRACT: Multivariate methods are overviewed from a systemic approach. The techniques are briefly discussed and simple examples are used to acquaint the nonstatistician with the rudimentary concepts and vocabulary. Comments on other possible modeling schemes will be briefly discussed. Some of the computer programs which are currently available are referenced.

KEY WORDS: sensory mechanisms, multivariate analysis, factor analysis, discriminant analysis, multidimensional scaling

Multivariate statistical techniques are rapidly becoming essential tools for the individuals working in the areas which, directly or indirectly, deal with the subjective-objective properties of consumer products. This thrust is manifest by the number of articles which are currently appearing in the literature, particularly the literature associated with the chemical, food, and tobacco industries. In these three industries, which share the tendency to accumulate enormous sets of data on their respective products, no doubt, past experience has demonstrated that research/development unguided by careful model building generally leads to ephemeral results.

The individual not normally trained in the specific statistical methodologies faces a severe barrier in attempting to utilize or interpret the outputs derived from these techniques. Consequently, a researcher placed in this position must resort to one of three courses of action: first, elect to spend one or more years of valuable time accumulating the pertinent literature and sorting through the myriad of details; or second, totally divorce one's self from the direct link with the experimental design/data analysis stages, and rely solely on the specialist; or third, spend a minimum amount of time by developing an overall acquaintance with the underlying principles and vocabulary.

Needless to say, the first alternative is quite time consuming and requires a great deal of self discipline which, in many cases, due to either industrial or academic constraints is simply not feasible or indeed possible.

The second alternative is in all probability the worst choice. Not only does the researcher place himself at a tremendous disadvantage in controlling the

[1]Group leader, Research and Development Department, Brown and Williamson Tobacco Corporation, Louisville, Ky. 40201.

flow required in completing the research mission, but in addition, will ultimately encounter or create a tremendous communications gap, tantamount to doom in today's age of accelerating technology. Given the evolving scenarios, the optimum choice is singularly the third alternative. The above arguments should not be misconstrued to imply that every researcher should become a quasi expert in multivariate methodology. Rather, what is suggested is that a basic knowledge be developed whereby one is aware of what techniques are available, what they can accomplish, and what their output represents in the framework of the experiment being conducted. This approach would not only serve as a time saving vehicle, but equally important, would establish and broaden the base of communications with the specialists.

The purpose of this paper is to construct an information bridge as inferred previously. Much of the arguments as well as the rudimentary examples are derived from informal courses and seminars conducted at one time or another for interdisciplinary groups at the Brown and Williamson Research Department. Some of the material has been presented in seminars on "market segmentation" sponsored by several universities. Additionally, some work which has been recently published or is in the preparation stage will be referenced. The integration of these is currently being developed and will shortly appear as a working text [*1*].[2]

The first part of the paper will discuss a "network" illustrating a classification scheme for the majority of the existing multivariate methods. Types of input data which are intimately connected with technique selection will be touched upon. This will be followed by a brief discussion of what the techniques are about without resorting to mathematical manipulations or arguments. In the second part, due to space limitations only two of the more widely applicable techniques will be illustrated. The examples used will be relatively simple and hopefully will demonstrate with some degree of substance the terms foreign to the nonstatistician.

Part three will consist of a commentary on some of the concepts which have evolved primarily from mathematical physics and information theory and how they may relate to the problem of subjective-objective measures. The reader should not regard these comments as dogmatic statements, but only as a forecast of what potentially is on the horizon in the field of subjective-objective analysis.

The last section will provide a compilation of computer programs which the reader may find useful as a minicatalogue. Additionally, a glossary and a bibliography are also included as a guide to the blossoming literature. This last part is, by no means, intended to be exhaustive.

Multivariate Methodology, A Pedestrian Approach

It has often been said that "... multivariate methods are generally tools which are efficacious and ..." Equally it has been often asked, "... nice,

[2]The italic numbers in brackets refer to the list of references appended to this paper.

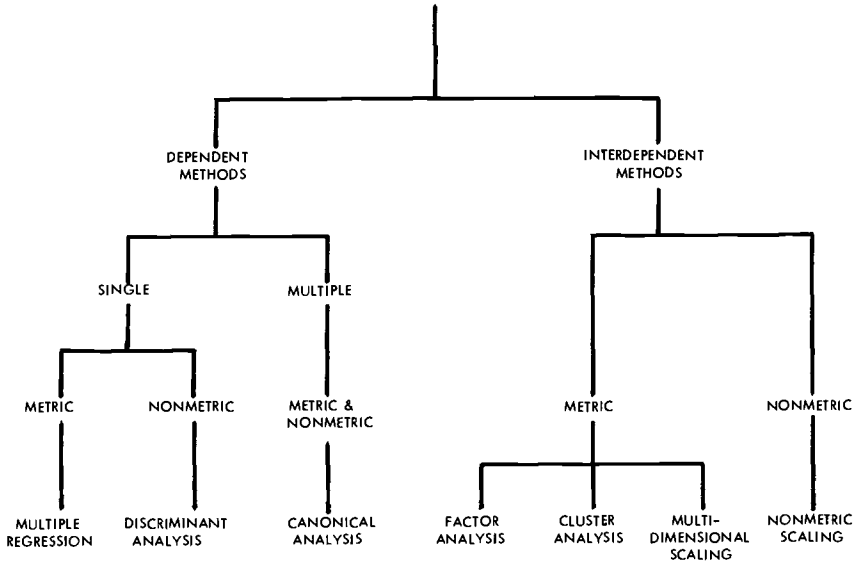

FIG. 1—*Systemic of multivariate techniques.*

neat, and all that but, when do I use them, what do I use them on, and what do I use?" In the first approximation, these are the types of statements made by the expert and questions asked by the would-be user. Fundamentally, the questions are very sound, and indeed should be asked, not only by the novice, but also by the expert before communicating to the nonexpert. In many instances, these questions may be answered by "laying out" a systemic of the critical decision points which would be normally encountered during the initial stages of experimental design (see Fig. 1).

Bearing in mind that we do not have at our disposal functional relations determining the properties of the system under scrutinization, a decision at the first level must be made. That is to say, is a dependent or independent model sought? In the former situation, we are generally being guided by the "in principle laws" which we think are operative. On the other hand, the latter situation reflects our state of ignorance about underlying laws and we must be satisfied with simply establishing whether or not the variates are interrelated.

Passing on to level two, we next attempt to ascertain the number of variates which should be treated as being independent or dependent. Again, this decision is guided primarily by intuition based on the assumed properties of the system, as opposed to computation.

The third level of the decision process concerns the basic properties of the data to be gathered. Specifically, are we dealing with variates which are expressed in terms of ratio scales, interval scales, or alternatively data derived from nominal or ordinal scales? Variates derived from interval or ratio scales are termed metric [2] since they express how much something is. Data

obtained from nominal or ordinal scales are referred to as being nonmetric [*3*] since they infer *which* rather than *how much* something is.

Thus, once having passed through the successive decision levels, the researcher is in a position to select the appropriate tool for data analysis. An additional benefit follows from the fact that the adoption of the systemic constrains an individual to perform a "gedanken experiment" relatively early in the game. This approach obviously should be exercised in any scientific approach, however, many of us quite often tend to forget the "holistic" aspects of problem solving and consequently require occasional steering.

On the dependent side of the systemic in Fig. 1, three general methodologies are available. Many variations of each technique exist; consequently, the three methods in reality represent families of techniques.

In a situation where a single metric dependent variable is to be related to a set of independent variables, multiple regression analysis should be used. The dependent variable is expressed in terms of a linear combination (LC) of the independent variables. The coefficients appearing in the LC are adjusted so that maximal collective correlation between the dependent variable and the set of independent variables is achieved. Examples in applying this technique to the problem of subjective-objective correlations have been previously discussed [*4*].

In many prediction problems, the dependent variable or predictant is split into several mutually exclusive groups. For example, in marketing problems consumers may be classified as being heavy, medium, or light users of a product. In subjective testing, one may be interested in dividing the responses derived from product descriptions into male and female or possibly age groupings. The "dependent variable" is thus expressed in nominal form and consequently is nonmetric. Problems of this type generally can be dealt with through the usage of multiple discriminant analysis (MDA). Multiple discriminant analysis is a technique which enables a classification of objects into two or more mutually exclusive categories via a set of independent variables. Specifically, the formal procedure requires maximization of the ratio of between group to within group variance with respect to the independent variables. A simple example in the next section will illustrate these points.

If a single dependent variable is measured under a varying set of conditions (treatments), establishing significance levels may be achieved by employing analysis of variance techniques. A logical extension of analysis of variance is required when several dependent variables occur and take on a range of values arising from various "treatments." For example, we may have several individuals who have provided subjective responses (in the form of metric data) on a particular product. One would (from experience) expect that a relatively large degree of correlation would exist between the variables. Because of the couplings, an analysis of variance would not suffice in testing the degree of response significance. Multivariate analysis of variance (MANOVA) techniques have been developed to deal with situations where

variable coupling exists. In this model the ratio of between group variance to within group variance is computed on the entire set of variables.

The method of canonical correlation can be formally viewed as an extension of multivariate analysis of regression. In this approach, one attempts to find a LC of one variable set (think of the independent set) that has maximal correlation with a LC of another set (the dependent set). In the strictest sense, the concept of independent and dependent variable is rendered somewhat irrelevant. Geometrically one should think of two spaces generated by the variable sets. Thus, one seeks to find a coupling between the two spaces. The attractive features of canonical correlation techniques are that both metric as well as nonmetric data can be dealt with, and conceptually different measures can be coupled. Needless to say, this in many instances, is "just what the doctor ordered" in the analysis of subjective-objective response.

The techniques appearing in the systemic on the dependent side, at least formally, strongly resemble one another. That this is so should not be surprising since they deal with both independent-dependent variables. These techniques are used because the researcher has an intuitive feel for an "in principle law." However, in the subjective-objective arena laws (in many cases) are equivalent to fantasies. Usually, before dependent methods are used, one must resort to some type of data sorting mechanism in order to develop a rough idea for the potential variate relationships.

The interdependent methodologies essentially provide vehicles for sorting through data sets. On the interdependent side of the systemic depicted in Fig. 1, four general families of techniques are shown. If the data are to assume metric form, factor analysis, cluster analysis, and multidimensional scaling should be used. These three particular methods are closely related and each has several variations.

Factor analysis [*5*] is a generic term for a multitude of techniques. Many individuals look upon this body of techniques with a degree of disfavor because of the lack of rigor. Nonetheless, when factor analytic techniques are used, in many cases, valuable insight is provided concerning variable interrelations.

Factor methods have two basic outputs: first, the original data are reduced, thereby summarizing the information contained in the observed or measured data set. Second, dimensions emerge which display variable couplings which are not obvious in terms of the language of the original variable set. It is at this point judgement must be exercised by the user. That is, the resulting new variables (called factors) must be interpreted. One then must look at each factor in terms of the old variables to make sure the analysis does not degenerate into an exercise in numerology. The factors represent a new coordinate system which may in turn be rotated [*6*] to optimally reveal symmetries contained within the data. Note that all the information is contained in the original data set and nothing new is introduced via factor methods. Factor techniques have been used in a number of consumer studies [*7-10*].

Several versions of factor analysis exist and very often confusion arises as to what exactly is being used (see glossary). The most common type of factor analysis is generally referred to as *R*-type. Adding to this confusion is the number of options available for rotating factors. The two general types of rotations which are used are called orthogonal and oblique [*5*]. All orthogonal rotations, irrespective of the particular computational algorithm, preserve the orthogonality of the factors. On the other hand, oblique rotation schemes allow for the possible coupling or correlations between factors.

A technique closely related to factor analysis is cluster analysis [*11,12*]. In this approach an attempt is made to group large numbers of objects (on the basis of their measured variables) into smaller mutually exclusive sets, such that the members of the individual sets have similar properties. For example, individuals can be clustered on the basis of how they perceive places or people [*13*]. Alternatively, products can be clustered via their respective attribute scores. Cluster techniques require a great deal of judgement on the part of the user. In many instances, it is not always clear what the cutoff point should be in forming a cluster or indeed how many clusters are desirable. This particular technique is generally used in connection with other multivariate methods [*13*].

In subjective evaluations on products, individuals or so-called experts attempt to describe product characteristics. Usually a set of descriptors are provided and judges rate the various product attributes. Alternatively, testing may also involve preference scores as well as an articulation of product characteristics. The data bank resulting from tests of this type are usually factor analyzed in order to obtain dimensions which characterize the product space as perceived by the respondents. In proceeding along these lines, several important questions come to mind. First and foremost, does the space in reality have any meaning? Are we, by virtue of testing, eliminating or introducing dimensions because of the constraints [*14*] imposed on the respondents?

Multidimensional scaling techniques attempt to skirt some of these potential difficulties. Recently, a great deal of attention has been given to these techniques [*9,15-18*]. Both metric and nonmetric multidimensional scaling deal with the analysis of similarity or preference judgments. In this way, the respondents are not constrained to provide detailed judgments concerning the test products. Scaling methods can be viewed as a type of "coarse-graining" approach in subjective evaluations.

In metric scaling the respondents provide their similarity or dissimilarity judgments in terms of either ratio or interval scales. Nonmetric analysis utilizes some type of ordering. The information in both approaches is transformed into distances and represented in a multidimensional space. Consequently, products then are classified by their relative interproduct distances. In most analyses, Euclidian spaces are generally used; however, other options [*18*], depending on need, can be exercised.

Illustrative Examples

As "advertised" earlier, this section will be devoted to illustrating some of the multivariate techniques. Two hypothetical problems will be used to acquaint the nonstatistician with both interdependent and dependent methods. The first example wil involve the use of factor analysis. Hopefully the reader will develop a better "feel" for the concepts and vocabulary used by the specialist. The second example will be nonnumerical and will emphasize the geometrical ideas of discriminant analysis.

Example I

We are given measures on five product attributes which are believed to be important for product positioning: taste, portability, flavor, compactness, and durability. From measures on these attributes, it is desirable to extract the structure with minimum dimension which describes the product.

Step 1—The data are standardized via

$$Z_i = \frac{X_i - \overline{X}_i}{\sigma_i} \tag{1}$$

where

X_i = i attribute,
$\overline{X}_i$ = mean value, and
σ_i = variance.

The new variable Z_i is dimensionless and has zero mean and unit variance.

Step 2—Using a representation in Z_i, the correlation matrix is calculated (see upper portion of Table 1). The correlation matrix is five by five and symmetric. That is, there are five rows and five columns. Interchanging the rows or columns, which is equivalent to reflecting the elements about the main diagonal leaves the matrix invariant. The off-diagonal elements contain information concerning the intervariable coupling, while the diagonal elements represent self-correlation terms.

Step 3—New variables are constructed by forming the appropriate linear combinations on the basis of the variable interrelationships. These new variables are related to the standardized variables through

$$F_k = \Sigma_j a_{kj} Z_j \tag{2}$$

where

F_k = k factor,
Z_j = j standardized variable, and
a_{kj} = factor loading.

TABLE 1—*Factor analysis, a hypothetical example.*

Taste	1.00	0.02	0.96	0.41	0.01
Portability	0.02	1.00	0.12	0.72	0.83
Flavor	0.96	0.12	1.00	0.56	0.09
Compactness	0.41	0.72	0.56	1.00	0.81
Durability	0.01	0.83	0.09	0.81	1.00

	Unrotated Factors			Orthogonal Rotation			Oblique Rotation	
Variable	F_1	F_2	h^2	F_1	F_2	h^2	F_1	F_2
1	0.58	0.81	0.99	0.02	0.99	0.99	0.13	0.99
2	0.77	-0.54	0.89	0.94	-0.01	0.89	0.93	0.09
3	0.67	0.73	0.98	0.14	0.98	0.98	0.25	0.99
4	0.93	-0.10	0.88	0.82	0.45	0.88	0.87	0.54
5	0.79	-0.56	0.94	0.97	0.01	0.94	0.96	0.10
Eigenvalues	2.87	1.80	4.67	2.52	2.15	4.67	...	...
Percent variance	57	35	93	50	43	93	$r_{F_1F_2}$	0.2

The factor loading is an estimate of the coupling between Z_j and F_k. It is important to note that no new information has been injected by the above steps.

Step 4—The factors are extracted in accordance with the variance that they "explain".... the first factor the greatest, the second factor the next greatest contribution, and so forth. Referring to Table 1 (see unrotated factors), we see that two factors are extracted. The percent variance associated with the first factors are extracted. The percent variance associated with the first factor is 57, the second 35 (values to the right of the decimal have been deleted). The total variance accounted for by the two factors is 93 percent. The first column lists the loadings of the individual variables on F_1, the second on F_2. Squaring the loadings and summing each column independently determines the eigenvalues. Each eigenvalue is an estimate of the variance associated with each factor. Alternatively, summing the square of the loadings across rows leads to the h^2 value or the communality. The communality is a measure of the variance of the particular standardized variable common to both factors. Thus, a large value of h^2 means that the combination of factors explains a large portion of the corresponding variable. A small value of h^2 implies little or none of the variance is explained by the factors taken collectively.

Step 5—The unrotated factor structure is not terribly enlightening with respect to interpretation. For this reason and only this reason, factor rotations are used. Two basic methods, with variations of each, are available to the analysts, orthogonal and oblique rotations. Orthogonal rotations preserve the information content in the unrotated structure. Consequently, this scheme can be looked upon as being equivalent to rotating a rigid body. This important feature is manifest by the fact that the h^2 values remain unaltered by orthogonal rotations, although the eigenvalues change. That this is so follows from the fact that the h^2 values are determined from summing the square of

the loadings for an individual variable across factors. The process of rotation simply redistributes the loadings on each factor. On the other hand, since the eigenvalues are determined from summing the square of the loadings over a factor, the redistribution process coming about by rotation should influence them. This is shown in Table 1. After performing the rotation, two dimensions emerge: variables two, four, and five are "heavily loaded" on F_1, while variables one and three are predominantly coupled to F_2. From this observation we can conclude that F_2 is a dimension dealing with organoleptic properties, and F_1 is a dimension which relates to the physical properties of the product.

As previously mentioned, several computational algorithms exist for each basic method of rotation. In the orthogonal approach, VARIMAX and QUARTIMAX methods are commonly used although other methods exist [*19*].

The results of oblique rotations on the factor structure are also tabulated in Table 1. Oblique methods differ from orthogonal techniques primarily through the allowance of factor-factor correlation. That is, the factors are no longer constrained to be uncorrelated. In this sense, oblique methods represent a more general class of transformations than orthogonal approaches. In most cases dealing with subjective-objective measures, it is sufficient to use orthogonal rotations. This opinion is based on the premise that the analyst is primarily attempting to determine the structure contained in the data set. The factors then can be used as either new variables for other multivariate computations (with the advantage of fewer variables to keep track of) or as a frame of reference for product descriptions.

In conclusion, two "rules of thumb" should be kept in mind when employing factor methodology: first, stop factoring when all factors with eigenvalues greater than unity are obtained; and second, rotate only factors with eigenvalues greater than one.

Example II

Consider the problem of predicting to which of two groups b (burley) or f (flue) samples of tobacco belong based on their measured chemical profiles [*20*]. For simplicity, let us assume only four variables are required to characterize the profiles. A graphical approach to the problem could be attempted by plotting all possible combinations of the (standardized) variables in a "scattergram." In Fig. 2 a scattergram is constructed using a representation defined by variables X_3 and X_4. It is worthwile to note that in an analysis of this type the identity of the variables need not be known. In fact, we could adopt such an approach for data condensation purposes before attempting time consuming identification work. The scattergram in Fig. 2 reveals little, if any, information concerning the problem. The means (denoted by b and f) are relatively close and there is a great deal of overlap of the individual observations. Plotting a scattergram in the X_1 and X_2 represen-

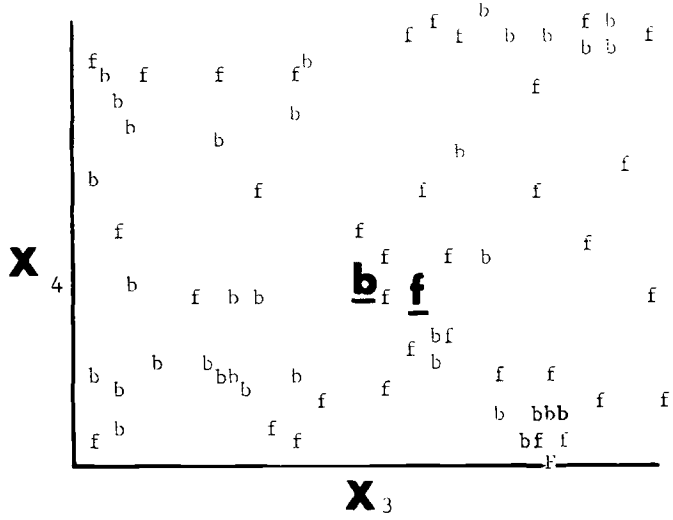

FIG. 2–*Scattergram of flue and burley tobacco chemical profile, poor discrimination.*

tation, however, does provide information. The means are well separated and very little overlap of the *b*'s and *f*'s occur.

The formal procedure for constructing a basis for classification of the observations as shown in Fig. 3 is provided by the method of discriminant analysis. One may look upon this technique as a process whereby the ratio of the spread between means, to the spread about means, is maximized [*21*]. There is more to the "business" than the preceding, however, which provides, in essence, a picture of what takes place. The variable which produces the largest ratio is the best discriminator. In this manner, a set of discriminating variables are constructed. The procedure produces a linear combination of variables called a discriminant function. Using the variable measures, the discriminant scores of each observation are computed. These scores are then used for classification of the samples; the line in Fig. 3 forms the boundary for classification, and the circle corresponds to a region from which a forecast of the classification probabilities are computed. In the case where three variables describe the discriminant function, a plane forms the classification boundary. By analogy, the many variable classification boundaries are determined by a hyperplane of appropriate dimensionality.

Discriminant analysis has been applied to several problems in the area of subjective-objective evaluations [*22*]. One of the more widely used approaches is the method of stepwise discriminant analysis [*22*], however, other computational methods have enjoyed a great deal of success and are described in the literature [*23*].

What Next

As reflected by the papers in this conference and also the literature, much effort is being expended in the area of subjective-objective evaluations. The

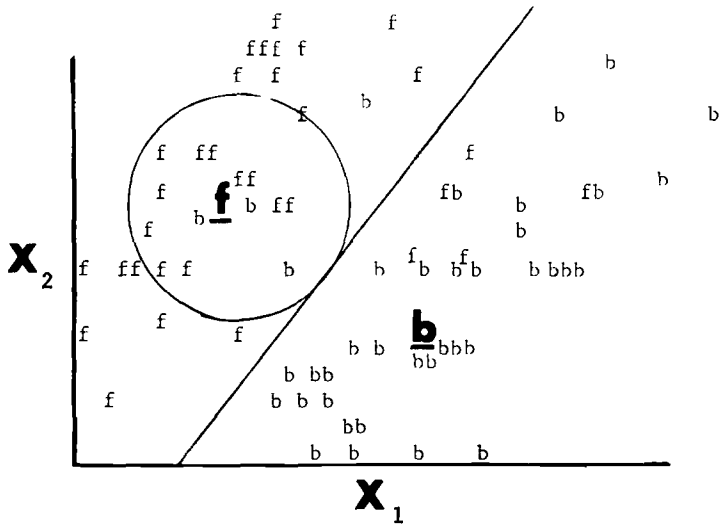

FIG. 3—*Scattergram of flue and burley tobacco chemical profiles, good discrimination.*

various statistical tools are being applied without too much difficulty because of the availability of computer systems. Needless to say, fundamental steps forward have been taken in correlating subjective-objective measures in recent years. However, what next? Should the community continue along the present course? Has time come to seek other avenues? If so, what directions?

Recently some of us [*24*] have become enamored by the possibilities of revisiting some of the models developed in the fields of mathematical physics and information theory. At this juncture, I do not know whether the effort would be productive. Nonetheless, on intuitive grounds, I feel it deserves some consideration.

One particular approach based on developments due to Shannon [*25*], although not directly applicable, may be with some modification important to subjective responses. Shannon recognized that all informational processes are selection processes. Central to the theory are the concepts governing the capacity of a channel for generating messages and the average amount of information per message per channel. In this respect, the analogy to some of the problems encountered in subjective testing is quite striking. Attempts have been made to formally apply this theory to problems in experimental psychology [*26*] and sociology [*27*] without a great deal of success. Some degree of success, interestingly enough, has been achieved by related formulations in the field of marketing [*28,29*]. It should be kept in mind that Shannon's theory was developed specifically to deal with telecommunications and formal application to other systems is dangerous. Perhaps the required spirit and concepts lie within this framework and await the proper modifications.

References

[1] Stungis, G. E. and Reynolds, M. L., *Industrial Applications of Multivariate Analysis*, Gordon and Breach, New York, in preparation.
[2] Moskowitz, H. R., *Food Technology*, Vol. 16, 1974, pp. 16-21.
[3] Neidell, L. A., *Journal of Marketing*, Vol. 33, 1969, pp. 37-43.
[4] Kramer, A., *Food Technology*, Vol. 23, 1969, pp. 66-68.
[5] Harman, H. H., *Modern Factor Analysis*, University of Chicago Press, 1967.
[6] Jennrich, R. I. and Sampson, P. F., *Psychometrika*, Vol. 31, No. 3, 1966, pp. 313-323.
[7] Sheth, J. N., *Journal of Advertising Research*, Vol. 10, 1970, pp. 8-17.
[8] Sheth, J. N., *Journal of Marketing*, Vol. 35, 1971, pp. 13-19.
[9] Frank, R. E. and Green, P. E., *Journal of Marketing Research*, Vol. 5, 1968, pp. 83-93.
[10] Frank, R. E. et al, *Journal of Marketing Research*, Vol. 2, 1965, pp. 250-258.
[11] Tryon, R. C. and Bailey, D. E., *Educational and Psychological Measurement*, Vol. 18, 1958, pp. 477-495.
[12] Bijnen, E. J., *Cluster Analysis*, Tilbury University Press, The Netherlands, 1973.
[13] Johnson, R. M., *Journal of Marketing Research*, Vol. 8, 1971, pp. 13-18.
[14] Rao, V. R. and Stungis, G. E., in preparation.
[15] Shepard, R. N., *Journal of Mathematical Psychology*, Vol. 3, 1966, pp. 237-315.
[16] Torgerson, W. S., *Psychometrika*, Vol. 30, 1965, pp. 379-393.
[17] Green, P. E. and Carmone, F. J., *Journal of Marketing Research*, Vol. 6, 1969, pp. 330-341.
[18] Green, P. E. and Rao, V. R., *Applied Multidimensional Scaling*, Holt, Rinehart and Winston, New York, 1972.
[19] Nie, N. et al, *SPSS*, McGraw-Hill, New York, 1970.
[20] Stungis, G. E. and Crain, W. O., 27th Tobacco Chemists' Research Conference, Winston-Salem, N.C., 1973.
[21] Sorenson, E. L. et al, "Users Manual for Statistical Computer Program Package," U.S. Air Force, Hanscom Field, Bedford, Mass., 1965.
[22] Powers, J. J. and Keith, E. S., *Journal of Food Science*, 1968, Vol. 33, p. 207.
[23] Young, L. L. et al. *Journal of Food Science*, Vol. 35, 1970, p. 219.
[24] SRF-University of Louisville, "Market Segmentation Techniques and Applications," Louisville, Ky., 1974.
[25] Shannon, C. E. and Weaver, W., *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, 1959.
[26] Cherry, C., *On Human Communication*, Science Editions, 1961.
[27] Buckley, W., *Sociology and Modern Systems Theory*, Prentice-Hall, New York, 1967.
[28] Herniter, J., *Journal of Marketing Research*, Vol. 10, 1973, pp. 361-375.
[29] The Hendry Corporation, "Hendrodynamics," Croton-on Hudson, N.Y., 1966.

Bibliography

Anderson, T. W., *Introduction to Multivariate Statistical Analysis*, Wiley, New York, 1958.
Cooley, W. W. and Lohnes, J., *Multivariate Procedures for the Behavioral Sciences*, Wiley, New York, 1962.
Engle, J. F. et al, *Market Segmentation Concepts and Applications*, Holt, Rinehart and Winston, New York, 1972.
Kendall, M. G., *A Course in Multivariate Analysis*, Charles Griffin, London, 1962.
Kruskal, J., *Psychometrika*, Vol. 29, 1964, pp. 1-27.
Van de Geer, J. P., *Multivariate Analysis for the Social Science*, W. H. Freeman, London, 1971.
Wilks, S. S., *Mathematical Statistics*, Princeton University Press, Princeton, N.J., 1943.
Williams, E. J., *Regression Analysis*, Wiley, New York, 1959.

Computer Programs

Levinsohne, Jay

Scalar–Multidimensional (nonmetric) scaling program, analyses symmetric unconditional dissimilarity data with no missing elements. The Euclidian model is used. Approximately 16 000 bytes of core can be used on IBM 370, 360, or 1130. Can be scaled up to 500 000 bytes.

Dobson, Rick

Torsca-9B–Nonmetric multidimensional scaling program. Written in neutral dialect of FORTRAN and is easily adapted to non-IBM machines. Approximately 1900 000 bytes of core

Young, Forrest

Polycon II–Nonmetric multidimensional scaling program. Handles symmetric and asymmetric matrices. Program convertible to non-IBM machines, approximately 210 000 bytes of core.

Carroll, J. D. and Chang, J. J.

Multidimensional scaling programs series, Bell Laboratories.

Clyde, D. and Cramer, E.

Multivariate statistical programmes, Biometric Laboratory, University of Miami.

Jones, K. J., Harvard

Multivariate statistical analyzer, 500K

Hemmerle, W. et al, University of Iowa

AARDVARK, statistical package

IBM

SSP, version III, broad collection of programs

Kendall, M. G., Scientific control systems

Cluster analysis

Service, Jolayne, North Carolina State University

SAS, variety of univariate and multivariate programs

Applebaum, M. and Bargmann, R. E., University of Georgia

Multivariate analysis of regular two-way classification

Biomedical Computer Programs, University of California, Los Angeles

Dixon, W. J., BMD, Editor

Correlation with transgeneration, 150K
Correlation with item deletion, 150K
Description of strata with histograms, 158K
Cross tabulation with variable stacking, 150K
Principle component analysis, 146K
Discriminant analysis, two groups, 124K
Discriminant analysis, multiple groups, 166K
Stepwise discriminant analysis, 163K
Factor analysis, 148K
Canonical correlation analysis, 114K
Multivariate analysis of variance, 160K

OSIRIS III

ISR, University of Michigan

OSIRIS III is a data management and analysis package. Package must be used with OS/360 or 370 computer system. Operating system must contain FORTRAN IV Level G, assembler and PL/I Level F compilers. Storage for package around 12 x 10^6 K.

General Purpose Programs–File copying, card copying, sorting and merging, etc.

Handling Multiple Punch Data–Multiple punch copy, merge check, frequencies, convert.

Transformation of Data–Index and recode data, summary description statistics, lag-lead, data simulation.

Correlation and Regression Analysis–Approximately seven programs including ordinal and categorical correlations.

Analysis of Variance–Variance analysis, one way, MANOVA.

Multivariate Analysis–AID 3–Generalized multivariate program based on analysis of variance techniques to explain dependent variable variance. Dependent variable must be continuous, independent variables may be ordinally or nominally scaled.

MCA–Examines the relationship between several variables and single dependent variable, then determines the effects of each predicator before and after adjustments because of intercorrelations.

THETA-AID–performs multivariate analysis of nominal dependent variables using nominal or ordinal scale independent variables.

Factor analysis and multidimensional scaling–variety of factor type programs as well as multidimensional scaling.

Cluster analysis–normal and hierarchical clustering programs.

Glossary of Technical Terms

canonical correlation, *n.*–a technique which is a generalization of correlational analysis. In this approach correlations are sought between two sets of data which may be conceptually different.

cluster analsyis, *n.*–a collection of techniques by which objects, people, attributes, etc. may be grouped together as logical entities on the basis of their similarities and differences.

dependent variable, *n.*–the variable which one wishes to explain as a function of other variables (called independent variables).

discriminant analysis, *n.*–a technique which allows the differentiation between two or more objects (groups) on the basis of the performed measurement.

factor analysis, *n.*–a collection of operations which may be performed on a set of data so that the degree of the data is reduced without losing relevant information. By virtue of the reduction of the data size, the techniques allow the recognition of certain nonobvious structures within the data set.

inner product of vectors, *n.*–a rule for multiplying two vectors. This type of multiplication results in a pure number (scalar).

linear transformation, *n.*–an operation on a vector, matrix, or a vector space such that the following rule is satisfied: $L(A) + L(B) = L(A + B)$; L being the transformation, A and B being either vectors, matrices, or vector spaces.

matrix, *n.*–a collection of numbers ordered in an array of rows and columns. The numbers within the matrix are called matrix elements. The rows or columns of the matrix are vectors; consequently, a matrix has the properties of a vector space.

***n*-dimensional vector space**, *n.*–a collection of n-vectors which obey certain algebraic rules.

nominal variable, *n.*–a variable whose values have no numeric significance, for example, automobile, sex, etc.

normalized variable, *n.*–a continuous variable which has been transformed so that it has zero mean and unit standard deviation.

***o*-analysis**, *n.*–when the data under examination are objects such as people, automobiles, bourbons, etc., the approach is called o-clustering (similar to q-factoring).

ordinal variable, *n.*—a variable whose values denote an ordering rather than possessing a definitive value; typically similarity data is represented in this manner.

***Q*-type factor analysis**, *n.*—factor procedure performed on associations between objects or individuals, for example, male/female, toothpastes, mouthwashes, etc.

***R*-type factor analysis**, *n.*—factor procedure performed on attributes, for example, personality variables.

regression analysis, *n.*—a technique which allows for the explanation of a given dependent variable on the basis of a set of observations through correlation methods.

scalar quantity, *n.*—simply a number, for example, 4 lb, \$3.50, or seven oranges.

square matrix, *n.*—a matrix which has the same number of rows and columns.

statistics, *n.*—quantities which are derived from a set of data and which summarize the data.

symmetric matrix, *n.*—matrix which remains the same after an interchange of rows and columns. All correlation matrices are of this class.

***V*-analysis**, *n.*—data under examination comprised of some mental abilities or some other attribute. The approach is referred to as *V*-clustering, similar to *R*-factoring.

variable, *n.*—a particular quantity which may assume a range of values, for example, Income I, which may have a range of definition from \$0 to \$5.

vector, *n.*—a quantity which possesses both magnitude and direction, for example, velocity as opposed to speed (former having both direction and magnitude, while the latter, simply a magnitude).

vector space, *n.*—a collection of vectors which obey certain algebraic rules. The usual cartesian coordinate system (two-dimensional) is a two dimensional space.

H. R. Moskowitz[1]

Applications of Multidimensional Scaling to the Psychological Evaluation of Odors

REFERENCE: Moskowitz, H. R., **"Applications of Multidimensional Scaling to the Psychological Evaluation of Odors,"** *Correlating Sensory Objective Measurements–New Methods for Answering Old Problems, ASTM STP 594*, American Society for Testing and Materials, 1976, pp. 97-110.

ABSTRACT: The use of multidimensional scales is outlined for two problems in odor perception. Multidimensional analyses of profile data allow the experimenter to obtain, at a glance, the necessary information about qualitative relations among odorants and their linguistic descriptors, as well as changes in the perception of the same odorant evaluated across time. All can be represented by the analysis procedures of multidimensional scaling.

KEY WORDS: sensory mechanisms, odors, taste, multidimensional scaling, geometry, multidimensional unfolding, dissimilarity

Recently, psychometricians have developed a new series of tools known as multidimensional scaling with which they are able to represent stimuli as points in a geometrical space of relatively low dimensionality. The history of these approaches is presented by Shepard [*1*],[2] and numerous applications of the method are detailed in a book by Green and Rao [*2*]. The rationale for representing the stimuli in a geometrical space is fairly straightforward. Those stimuli that are subjectively similar (either in terms of a specific attribute or else in their overall impressions) are placed close together in this space, whereas stimuli that are subjectively dissimilar in quality are placed far away. The scientist can obtain a quick impression of the qualitative variations among stimuli by inspecting the overall geometry of the space (namely, the relative placement of points) and can derive by various methods clusters of subjectively similar stimuli. Perhaps the most important *present* use of the methods classified under multidimensional scaling is representation of stimulus relations. The procedure does not reveal information that is truly "new," but rather presents the data in a form which allows the experimenter to obtain insights into relations among stimuli (and in some cases relations among individual observers).

[1] Head, Sensory Psychology and Methodology Unit, Department of the Army, U.S. Army Natick Development Center, Natick, Mass. 01760; presently with MPI Sensory Testing, N. Y., N. Y. 10021.

[2] The italic numbers in brackets refer to the list of references appended to this paper.

In order to envision how such spaces are developed, it is instructive to consider a simple analogy. Suppose the configuration of the United States is unknown, but a large number of intercity distances are known. For example, if 100 cities of the United States are to be located on the map, then initially one is provided with all 100(99)/2 or 4950 intercity distances, but is not informed where these cities lie relative to each other. Such information begins to constrain the *relative* locations of the cities. Chicago could not be located next to New York, for that would violate the distances between Chicago and Los Angeles and between New York and Los Angeles. As the number of such intercity distances diminishes, the location of the various cities on a map becomes decreasingly fixed and more subject to error in localization. With the full set of constraints operating, the map of the United States will emerge, but its positioning relative to the North-South and East-West axes might be inverted or rotated any number of degrees.

If individuals are instructed to rate the dissimilarity among pairs of stimuli, then these dissimilarity values can be treated as if they represented interpoint distances. A variety of computer programs have been published which will process these interpoint distances to yield a "map" in which the various stimuli are embedded.[3]

Use in Odor and Taste Studies

If observers are instructed to rate the qualitative dissimilarity among pairs of odorants or pairs of taste substances, then these odorants or taste substances can be embedded in geometrical spaces, usually of low dimensionality (for example, one, two, or three dimensions). Woskow [*3*] reported studies in which observers rated the dissimilarity among 25 odorants (chemicals of known composition) and reported that the first dimension of the odor map was hedonics (pleasantness/unpleasantness). Since then, most of the studies (for example, see Refs *4*, *5*, and *6*) have concurred that the initial dimension to be extracted is that of hedonics. The remaining dimensions in odor vary so much with the choice of odorants that little in the way of generality can be extracted from the various reports. Yoshida [*7*] has presented a variety of data sets for olfactory stimuli, analyzed by several methods, and suggested that there is a large number of dimensions along which odorants vary. For taste, studies by Schiffman and Erickson [*8*] suggest at least three different dimensions, including hedonics (perhaps as the most important one).

As can be inferred from the foregoing discussion, the major aim of researchers using multidimensional scaling in the chemical senses (and of those using the methods for other stimuli) is to uncover the basic, fundamental, underlying dimensions which people use in differentiating stimuli qualitatively. The lack of concensus about the underlying dimensions may never be resolved if experimenters continue to widely use different assortments of chemicals and

[3]P. Green and V. Rao have detailed several of these programs, that are available from Bell Telephone Laboratories, and illustrated their use in data processing.

if they rely solely upon the maps which are derived from the analysis of small sets of different chemical compounds.

Another use of multidimensional scaling in the analysis of odor percepts is to portray individual differences. Individuals show characteristic differences in the way that they describe odors, partly due to the fact that odors have the power to evoke memories idiosyncratic of each individual. In addition, chemical odors may provoke ambiguous impressions, since they are not necessarily representative of ecologically valid, real-world stimuli. In one set of experiments in which observers judged the overall dissimilarity of 15 chemicals (for example, xylene, isobutyl isobutyrate, etc.), individuals differed quite dramatically in the way that they rated pair-wise dissimilarity. Odorants highly dissimilar to one individual were moderately similar to another and, in some instances, quite similar to a third. This study of 15 individuals was reported, for grouped data, by Moskowitz and Gerbers [*5*] and an analysis of individual data was reported by Moskowitz and Pell [*9*]. In a second study, this time with 20 food aromas procured from flavoring houses and diluted to approximate the actual odor of foods, the individual variations in difference judgments were substantially smaller [*10*].

The basic tool for uncovering individual differences by means of multidimensional scaling is the INDSCAL approach (individual differences multidimensional scaling). Carroll and Chang [*11*] have developed a computer program by the same name which embodies the hypothesis that: (*a*) all individuals share a common, grand, or public geometrical space; and (*b*) each individual possesses a private set of weighting factors which may shrink some dimensions and expand others. The INDSCAL computer program permits the experimenter to analyze sets of dissimilarity matrices. The input is a set of such matrices, one matrix per individual. The output is a "grand" geometrical space, containing n dimensions, in which the stimuli are embedded and a series of weighting factors, one per dimension per observer, represent the expansion or shrinkage factors.

Multidimensional Scaling with Profile Data

The traditional analysis of pair-wise dissimilarities data produces an unwieldy number of comparisons. For n stimuli the total number of comparisons required is $n(n-1)/2$ or for large values of n, a number of comparisons which varies with $n^2/2$ (minus a small amount). For 10 stimuli the number of comparisons jumps to $10(9)/2 = 45$. Rarely can more than 25 stimuli be evaluated in this approach in a reasonable period of time because of the tedious nature of the comparisons and their large number. Consequently, rarely are odor maps (or taste maps) generated with more than 25 to 30 points.

In order to circumvent this problem about limitations in the number of stimuli, experimenters have attempted to embed both the descriptor terms and the stimuli in the same geometrical space through the method of "multi-

dimensional unfolding." Coombs [*12*] discussed the theory of unfolding with which observers' judgments about the dissimilarity (or similarity) of a stimulus to a concept sufficed to locate the two of them in a conjoint geometrical space. In the present case, profile tables can be used as the input to the unfolding analysis.

In the traditional profiling method, the observer is provided with a series of descriptor terms which apply to the stimulus. In addition, a scale is set up so that gradations in profile values can be established. In some instances the observer may be asked to estimate how strong a stimulus odorant appears to be on several odor attributes (harshness, fruitiness, spiciness, putridity, etc.). The scale that is used may be either a category scale (for example, 0 = not at all applicable, 1 = slightly strong, 9 = extremely strong) or a ratio scale (0 = not at all applicable, 10 = moderately strong, 200 = 20 X as strong as the value for 10). The category scale [*13*] is similar to the centigrade scale for temperature; differences in category scale values reflect differences in strength. There is no fixed zero point for the category scale, nor can ratios be inferred (so a category value of 8 does not mean 2 times as much as a value of 4). The ratio scale does permit assessments which have these proportions or percentages. Ratio scales were extensively used in the direct assessment of sensory intensity by Stevens and his co-workers during the period starting in 1953 [*14*]. As will be discussed, ratio scales have also been used for the evaluation of qualitative dissimilarity between pairs of odorants.

In its practical application for the evaluation of odorants, the method of unfolding has myriad uses. The stimulus set need not be limited to 25 or so odorants, but can encompass several hundred if the experimenter so desires. Nor need the experimenter limit the number of descriptor terms. Twenty, fifty, or even 100 may be used, as long as the observer is provided with stimuli in short experimental sessions and is "recalibrated" to use the same scale every time. Since the observer profiles a single odorant stimulus at one time, and since the meaning of the profile descriptor terms can be explained (and even illustrated by appeal to cognate terms or to representative stimuli), the data thus generated can be made as error free and as bias free as possible through careful experimental control.

As noted before, the unfolding method embeds both the stimulus odorant and the descriptor term in the same geometrical space. The experimenter is thus able to visually determine which odorants lie on the line connecting pairs of descriptors. These intermediates are those odorants which exhibit both descriptor notes, but to different degrees. If odorant X lies closer to one descriptor (P) than to another (Q), but nonetheless lies on the line connecting them, then the odor quality of X is intermediate, but more similar to P than to Q. Ideally, an entire series of chemicals could be sought which occupy the line between descriptors P and Q, but lie on different parts of the line. These chemicals, $X_1 \ldots X_n$ would represent a graded series of intermediates, which exhibit variations in similarity. These variations (relative to descriptors

P and Q) are quantified by the distance of each descriptor from P and from Q. That distance is provided by the standard formula for Euclidean distance.

Geometrical configurations obtained by embedding both stimuli and descriptors are provided by a variety of programs. The one used here is known as M-D-SCAL 5M [*15*]. The input data is a profile matrix in which the data comprises m columns (corresponding to m different descriptor terms) and n rows (corresponding to n different stimuli). The higher the profile entry in cell $m_i n_j$, the more "similar" stimulus n_j is to descriptor term m_i. Conversely, the lower the profile entry in cell $m_i n_j$, the more "dissimilar" the stimulus is to the descriptor term. The output of the analysis is a geometrical map jointly embedding the m descriptor terms and the n stimuli.

Study I, Aroma of Carrot Root Essential Oils

Figure 1 shows the results of a study in which observers were instructed to evaluate the appropriateness of 52 descriptor words as applied to 36 chemicals and concepts of chemicals. A full version of the experimental procedure appears in Albran, Moskowitz, and Mabrouk [*16*]. Briefly, the observers were instructed to sniff small bottles containing the chemicals (sonicated to emulsify them in water) and then to rate how appropriate each of the

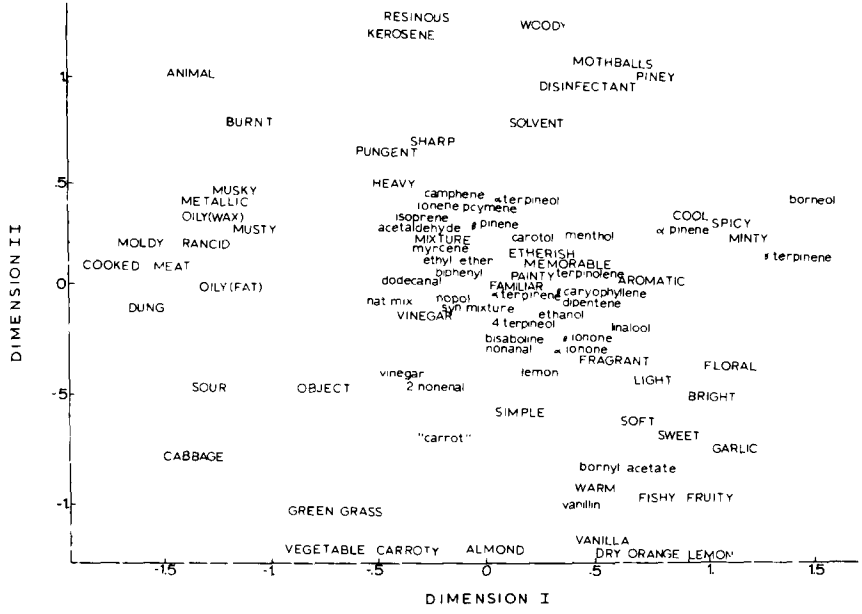

FIG. 1—*Two dimensional representation of the odor qualities of 36 odorants evaluated on 52 attributes, by a category scaling procedure. The odorant chemicals are listed in small letters, whereas the concepts (descriptor terms) are capitalized (data adapted from Ref* 16). *The data are the geometrical representation, in unrotated coordinates, obtained via the M-D SCAL program.*

descriptor words seemed to them to be. The chemicals were presented to the observers in small, wide-mouthed scintillation bottles. As is done in all studies of this type, where order bias and fatigue are operative factors, the order of samples was varied from one observer to the next, and numerous rest periods were interspersed in the experimental period. The observers had had previous experience in evaluation of odor qualities by this sniff-bottle procedure. The observers were free to pace themselves, and no time constraints were placed on them. It usually took several minutes to rate each odorant using all 52 descriptors, a time factor which militated against olfactory fatigue.

The scale used in the experiment was a 0 to 5 category scale (0 = not at all applicable, 5 = extremely applicable). Note that applicability here is not the same as intensity. Rather, the emphasis was upon appropriateness of the descriptor to the stimulus. However, the profile matrix which resulted was treated as if the entries reflected dissimilarities (0 = maximally dissimilar, 5 = maximally similar).

In Fig. 1 the odorant chemicals are shown as words in small letters, whereas the concepts are shown as capitalized words. One of the odorants is "carrot." This was not an actual odor, but rather a bottle labelled carrot, which was assumed to represent the panelist's concept of what a carrot odor should be. That is, for this *conceptual stimulus* the panelist was instructed to treat it *as if* it were an odorant and to scale it accordingly, using the various descriptor terms. The stimulus odorant marked "nat mix" was an extract of carrot root, while the stimulus marked "syn mix" was a mixture of the various aroma chemicals typically found in carrot root extract.

The array of 52 descriptors and 36 chemical stimuli was subjected to analysis with the M-D-SCAL 5M program [*15*], using the "unfolding option" which places the row and column stimuli into a joint geometrical space, based upon the dissimilarity (or distance) values in the matrix. The computer program was run several times to obtain several different solutions (or geometrical configurations) simultaneously embedding the chemicals and the descriptors. These solutions were all two or three dimensional. Finally, the solution with the lowest "stress value" (a measure of badness of fit, equivalent to the root-mean-square deviation between predicted and obtained distance values) was selected for two dimensions, as shown in Fig. 1. The distance between the descriptor and the chemical stimulus is a measure of the degree of appropriateness of the descriptor term for the stimulus. Since the M-D-SCAL 5M program attempts to achieve a single, global configuration embedding all stimulus odorants and descriptors, some specific distances were not adequately represented in the resulting map. However, as a general map, the configuration in Fig. 1 is adequate, based upon the low stress value achieved.

In Fig. 1, large distances mean that the descriptor term is inappropriate for the stimulus chemical, whereas small distances represent the opposite, namely, that the descriptor is appropriate. For example, the descriptors fragrant, floral, light, and bright are appropriate for the chemical linalool, but inappropriate

for the chemical dodecanal or borneol. An example of the failure of such mapping to fully reproduce the data matrix can be seen by the placement of vanillin, which is close to the concept of vanilla, dry, fruity, fishy, and lemon. All of those except the concept of fishy are appropriate. However, because the computer program was forced to locate the stimuli in two dimensions, some unusual proximities resulted. Minimization of "badness-of-fit" globally produced these unusual proximities, which would be eliminated (presumably) if geometric spaces of three, four, or higher dimensionality were recovered.

In order to make sense out of the rules which might lead to the placement in two dimensions, the experimenter need only consider which chemicals lie at the extremes (or else which descriptors lie at the extremes when multidimensional unfolding is used). Here the first dimension (I) comprises cooked meat, moldy, rancid, oily, etc. at one end, and the concepts of cool, spicy, minty, floral, and bright at the other (garlic, fishy, and such seem not to be appropriately placed in this two dimensional configuration). Hence, Dimension I is probably food-like (heavy, animal, organic) aromas at one end versus estery, sweeter, and lighter aromas at the other end. Relatively few chemicals congregate around the lower portion of Dimension I, presumably because most of the essential oils and chemicals found in carrot root aroma are sweet, estery, but rarely animal-like. The second dimension (II) varies from vegetable, almond, lemon, and orange at the low end to woody, resinous, kerosene, mothballs, piney, and disinfectant at the high end. Dimension II, therefore, seems to differentiate between food-like aromas and chemical-like aromas. Again, there are misplacements, animal and burnt belong at the low end of the dimension, but this misplacement probably would be rectified in a three dimensional map.

Several things should be borne in mind when evaluating data of the type shown in Fig. 1.

1. The unfolding procedure generates all distances between odorant chemicals and all distances between descriptor terms. Such distances were not obtained experimentally, in the actual study, but were obtained from the placement of chemicals and descriptors in the joint space.

2. If the entire pair-wise set of comparisons were to be made, as is the traditional experimental design evaluating dissimilarities, then the profile maxtix would have to be augmented from a 52×36 matrix, to a $(52 + 36) \times (52 + 35)/2$, in which each element would be compared against every other element. The former requires 1872 entries for each chemical to be compared to each descriptor. The latter requires 3828 entries. The savings by use of the profiling matrix is on the order of 50 percent fewer judgments required by the panelists.

3. Because odor descriptors and odor chemicals are simultaneously embedded in a joint space, odor nuances can be more easily seen via the geometrical map. That is, along a line connecting two odor descriptors (for example, cool and etherish) there will lie a series of chemicals which share in common some degree of coolness and some degree of etherishness. Here the

chemical pinene has these two properties, but it is cooler versus being etherish. For the chemicals lying between the descriptors fragrant at one end and heavy at the other, there lie a whole series of compounds. The ionones are more fragrant than they are heavy; the chemical terpineol is approximately midway between the two descriptors, whereas ionene and camphene are far more heavy than they are fragrant.

4. Many of the descriptor terms are located towards the outside of the geometrical space, whereas many of the chemicals are located towards the inside. Moskowitz and Gerbers [*5*] noted this similar placement of chemicals towards the inside for evaluations of 15 odorants and 15 chemical descriptors, for both direct estimates of dissimilarity and for multidimensional unfolding, as done here, respectively. The fact that descriptor terms tend to lie towards the extremes means that these terms are more dissimilar to each other than chemicals are to each other, respectively. Apparently, nothing is as "minty"as one's concept of ideal "minty", nor as putrid as one's concept of ideal "putrid."

5. An actual aroma, and even a synthetic mixture made up to resemble that natural aroma, may not typify the ideal aroma quality. Here, the concept of carrot differs from the aroma of the carrot root oil. This means that for some (if not all) aromas, the concepts of the aromas differ from examples of the aroma. Whether this is due to variations in the actual aroma itself (due to varying concentrations of essential oils) or whether there are true differences perceptually and cognitively between concepts of aromas and aromas themselves waits for further experimentation. Moskowitz [*17*] using the method of proximities analysis (multidimensional scaling of dissimilarities between pairs of aromas, pairs of aromas descriptors, heterogeneous pairs of aromas versus aroma descriptors) noted that for some aromas (for example, lemon and cantaloupe) the concepts of the aroma and the aroma itself could be superimposed in a geometrical aroma space. For other aromas, such as sausage, there was a difference between the observer's concept of the aroma of sausage and the actual sausage aroma presented to her.

6. Small differences in aroma concepts can be elucidated by this procedure. For example, the concepts of waxy oily and fatty oily occupy different positions in the geometrical space. This differentiation comes about because the two terms are used differently to describe the same set of odorant stimuli.

7. Since the geometrical map represents imputed dissimilarities between various stimulus chemicals and descriptors, the distances on the map represent relative dissimilarity values. Were the experimenter so inclined, he could determine iso-dissimilarity contours in the space by drawing circles from a center point of predetermined radius. All descriptors and chemicals which intercept that circle are, by definition, equally dissimilar to the central point.

Studies of Aroma Perception Over Time

Another use of the multidimensional scaling procedures is to represent how a group of individuals perceive qualitative dissimilarities over time when these

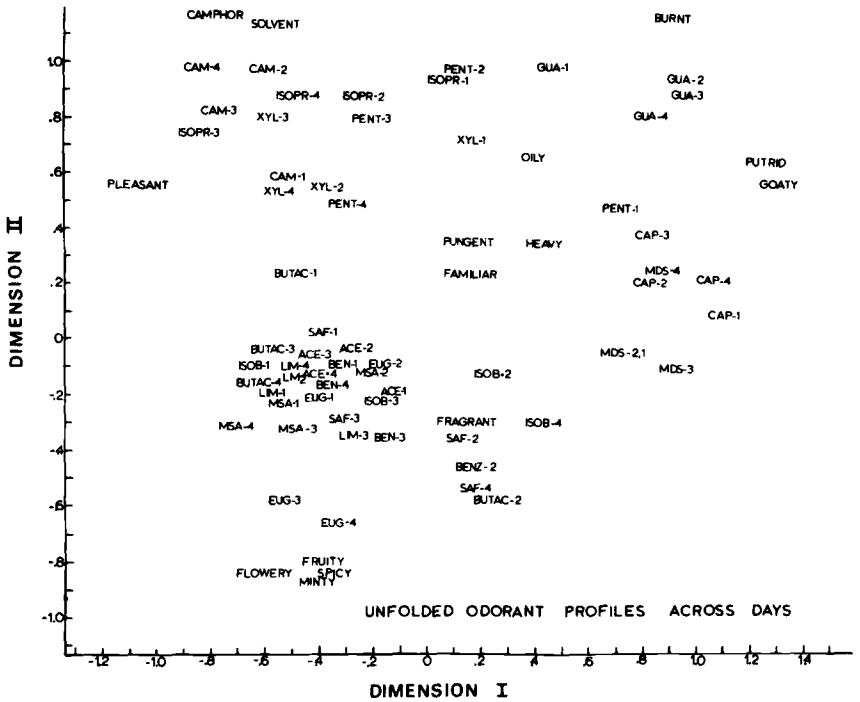

FIG. 2—*Two dimensional representation of the odor qualities of 15 odorants, evaluated on 17 descriptors. Each odorant was evaluated on four separate days by 15 individuals. The mean profile ratings for each day were unfolded into a two dimensional geometrical space. Numbers at the end of each chemical indicate whether the chemical was evaluated on Day 1, 2, 3, or 4, respectively.*

odorants are presented over a course of several sessions. If the observer tends to focus in upon new attributes with repeated experience, and thus adopts another criterion for making dissimilarity estimates, then this shift in criterion will show up as a movement of the odorant position in the geometrical space, assuming, of course, that the same odorant is presented in the geometrical space as different points (corresponding to the different presentations). On the other hand, if the observers perceive the odorants in the same way, day after day, and use the same scale to represent dissimilarity (or if the same profiling procedure is used day after day), then the position of the odorants should be invariant in the geometrical map.

Figure 2 presents the results of an experiment in which the observers profiled the odor of 15 chemicals. Each day the observers rated the degree to which 17 attributes applied to the 15 odorants. The procedure for profiling was relatively straightforward. During a 4-hour session, the observer rated the same 15 odorants four times (total = 60 samples, 4/odorant). Each time an odorant was presented, the observer rated that odorant on a magnitude estimation scale for the 17 descriptor terms. The ratings were averaged across

the 15 observers for each day to yield a matrix of 15 (odors) X 17 (descriptors). Four such matrices were obtained, one for each of the four days.

The chemicals evaluated in this second study were camphor (CAM), xylene (XYL), pentanol (PEN), caproic acid (CAP), guaiacol (GUA), methyl disulfide (MDS), eugenol (EUG), butyl acetate (BUTAC), isobutyl isobutyrate (ISOB), safrole (SAF), limonene (LIM), benzaldehyde (BEN), acetone (ACE), isopropanol (ISOPR), and methyl salicylate (MSA).

The descriptor terms used in the profiling were flowery, fruity, spicy, minty, pungent, familiar, pleasantness, goaty, putrid, fragrant, oily, burnt, solvent-like, camphor, heavy, and total odor intensity. Total odor intensity was not used in the analysis of the profile entries.

A full description of the procedure has been previously presented [5]. Briefly, the odorants were initially selected to be dissimilar qualitatively. They were matched by the experimenter to equate for odor intensity and were presented to the observer in small scintillation bottles. Each odorant was saturated onto a cotton ball, and then this ball was inserted into the scintillation bottle. The large surface provided by the cotton ball permitted evaporation to occur quickly and, thus, allowed for a maintained steady state equilibrium cloud of vapor within the bottle. Although sniff bottle procedures may not provide sufficiently accurate control of concentrations, they are quite useful in the evaluation of odor quality where changes in the exact concentration of the vapor phase do not play a dramatic role (if minimized).

The M-D-SCAL 5M program was again used to derive a geometrical map of each odorant and odor descriptor, in a geometrical space of two dimensions. The input matrix comprised a rectangular matrix of 16 rows (corresponding to the 16 descriptor terms) and 60 columns (corresponding to the 15 odorants evaluated on Day 1, Day 2, Day 3, and Day 4, respectively). That is, the computer program was not instructed to consider all odorants as replicated on four days, but rather treated each odorant evaluated on a new day as an entirely separate odorant. The two dimensional output map, shown in Fig. 2, clearly shows the distinction among days. Each odorant has a number appended to it indicating which day it was evaluated. It should be borne in mind that this unfolding procedure works only if the scales used by the observers are consistent from one day to the next, a finding substantiated by appeal to ratings for isobutyl isobutyrate which were virtually unchanged across four days of testing [5].

Figure 2 shows the two dimensional representation. The important findings from this analysis of the odor profiling data are the following.

1. Odorants and their descriptors tend to congregate in clusters. The lower left cluster comprises those odorants that are primarily pleasant, as well as those descriptors that are positive or hedonically pleasant. Safrole (odor of sassafrass oil), limonene (odor of lime), eugenol (oil of cloves), methyl salicylate (oil of wintergreen), acetone (solvent, sweet in high dilution), butyl acetate (banana oil), and the descriptors fruity, flowery, spicy, and minty all cluster in this region.

Two other clusters appear. Those at the top right, representing harshness and putridity, as well as burnt all congregate in that region. The cluster at the top right comprises odorants which are solvent-like, including isopropanol, camphor, xylene, and pentanol.

2. Odor descriptors again lie at the outer part of the space, whereas the odorants themselves lie closer to the center. Descriptors lying closer to the middle are only those such as pungent, heavy and familiar. These descriptors apply to many of the odorants (namely, they are "general descriptors"), and by virtue of such overall applicability these descriptors are forced to lie in the middle, near all of the odorants. Other descriptors, such as camphor, solvent, and burnt lie at the top, whereas the cluster of more positive hedonic descriptors (for example, fruity, spicy) lie at the lower left.

3. The actual coordinates of points on the dimensions makes little difference in terms of the meaning of the geometric representation. The multidimensional scaling concerns the interstimulus distances. The map shown in Fig. 2 may be rotated any number of degrees or completely turned inside out. Additive constants can be added to all ordinate and abscissa values, just as long as the distances themselves remain invariant.

4. The position of many of the odorants does not dramatically change from one day to the next in Fig. 2. Rarely does an odorant shift its location entirely, crossing from one section of the map to the opposite side. Such invariance in the position of points in the space is a powerful argument for the reliability of the observer's judgments with the profiling procedure. Were the program to yield random configurations with little or no meaning or were the observers to assign profile values indiscriminately (or even with a substantial amount of error), then one might expect the position of an odorant judged on Day 1 to be quite different from the position of the same odorant judged on Day 2, Day 3, and Day 4. By and large, the positions are fairly close to each other. This must mean that: (*a*) the observers use the scales (magnitude estimation of degree of descriptor applied to the odorant) in reliable ways; (*b*) they maintain the same criterion for judgment across days; and (*c*) in some specific cases odorants move to the corners of the geometrical space (although the movement is not usually large, for example, methyl salicylate, camphor). This can be interpreted as meaning that the observer applies *fewer* descriptors to these odorants and concentrates pretty much on the primary descriptor term.

5. If the odorants are assumed to remain unchanged in their position across days, then another interpretation of the movement can be made. This is that Fig. 2 really represents a projection onto two dimensions of a figure which is really three dimensional. Points close together on this three dimensional surface may be forced apart when the projection is made. If such an interpretation is accepted, then one possible candidate for the original configuration is a football shaped figure, whose major axis lies at a 45 deg angle with the two orthogonal axes. The tip of this football would be the descriptor burnt, whereas the other tip would be the four descriptors, flowery,

fruity, spicy, and minty. In the middle lies camphor (at the extreme left) and methyl disulfide (at the extreme right), two examples of relatively unpleasant odorants. Camphor smells like a chemical and is nonfood-like, whereas methyl disulfide smells rotten and far more organic. In the middle of this football figure would lie such ambiguous odors such as xylene and pentanol.

Correlations of Profiling Data with Instrumental Data

Two decades ago the English odor scientist, Moncrieff, suggested that a profitable way of analyzing the underlying mechanisms of odor perception might be to obtain for each odorant an objective "signature." This signature may be obtained via a collection of physical measures on these odorants. Moncrieff's original suggestion [*18*] was to relate odorants on the basis of the ability to absorb onto a variety of receptive surfaces, such as silica gel, fats, and fuller's earth. Concomitant work on the development of subjective signatures by profiling would provide a complementary set of relations among the same odorants. The outcomes of these procedures, one instrumental and the other subjective, are two profiles whose concordances may be established. Eventually the similarities and dissimilarities would be clues as to the physical similarities which correlate with subjective olfactory similarities.

This is the beginnings of what might be called a "grand scheme" to unlock the mechanism of odor perception. It is still around today in various forms. Amoore and Venstrom [*19*] correlated molecular shape with olfactory properties, and for each of more than 100 odorants they obtained two profiles. One of these was an instrumental measurement profile, comprising the degree of overlap of the shape of each test odorant onto the molecular shape of each of five specific primary odorants. The second was a subjective profile, comprising estimates of the similarity between the test odorant and each of seven standard odorants, which represented exemplars of Amoore's seven standards (floral, minty, pungent, putrid, camphoraceous, ethereal, and musky). Similar studies by Schiffman [*6*] concerned the relation between the geometrical spaced induced by similar estimates of odor quality (obtained via subjective scaling) and the physico-chemical parameters of odor stimuli (for example, chemical structure, functional groups, physical constants associated with the molecule). Schiffman's approach was to regenerate the odor quality space, obtained first by subjective estimates, by means of weighted measures on physical variables and attributes.

This scheme is an attractive one, but it is just the first step in the potential use of multidimensional scaling to bridge the gap between subjective perceptions of odor and physical characteristics of odorants. The use of such scaling is to portray the picture of the olfactory world, whether that world be generated via subjective estimates of dissimilarity among pairs of odorants (and thus the laws developed pertain to the mental "algebra") or generated via relations among the physical attributes of odorants.

Multidimensional Scaling and Models of Olfaction

Beyond the stage of representing similarities among chemicals and modelling physico-chemical relations by means of geometrical representation, there lies an entirely new field of research to which the scaling methods of multidimensional scaling can be applied. This field is to aid the scientist formulate questions about how olfactory perception may work. The multidimensional representation of odorants need not be confined to a variety of chemically dissimilar stimuli in which it is left up to the ingenuity of the scientist to determine what are the criteria which the individual employs in making judgments of qualitative dissimilarity. Rather, if chemicals and their mixtures are represented in the space, a variety of questions may be asked with appeal to extraneous information (for example, physico-chemical parameters) which may correlate only incidentally with olfactory quality. Some of these questions are as follows.

1. In a mixture of odorants, do the components comprise a convex region in the geometrical space, and if they do, do the odorant mixtures in turn lie in this convex region? That is, is the quality of the mixture somehow intermediate between the qualities of the components? If it is, then the prospects are fair for an algebra of odor mixtures. If not, then the prospects for a true algebra of odor quality perception must wait for a more fertile approach, since knowledge of component odor qualities does not permit the experimenter to make statements about the quality of the mixture.

2. Can the odor space be used to identify perceptual anomalies, much as multidimensional scaling of color similarities reveals two quite distinct color spaces, one for normals and the other for color blind people? Wish and Carroll [*20*] reported the use of the INDSCAL program to generate the grand geometrical space for color perception and the individual weighting factors corresponding to normals and to color blind individuals. Work by Amoore [*21*] on specific anosmia suggests a variety of such anosmias. Amoore's work concerns primarily the variation in odor threshold as the criterion for specific anosmia, without undue concern for variations in odor quality.The use of multidimentional scaling may illuminate the different odor worlds that anosmia reside in and point the way to still further fruitful research in odor mechanisms.

References

[*1*] Shephard, R. N.,*Psychometrika*, Vol. 39, 1974, pp. 373-422.
[*2*] Green, P. and Rao, V., *Applied Multidimensional Scaling*, Holt, Rinehart, and Winston, New York, 1972.
[*3*] Woskow, M. H. in *Theories of Odor and Odor Measurement*, N. Tanyolac, Ed., Robert College, Turkey, 1968, pp. 147-188.
[*4*] Berglund, B., Berglund, U., Engen, T., and Ekman, G., *Scandinavian Journal of Psychology*, Vol. 14, 1973, pp. 131-137.
[*5*] Moskowitz, H. R. and Gerbers, C., *Odors : Evaluation, Utilization, and Control*, Annals of the New York Academy of Sciences, Vol. 237, 1974, pp. 3-16.

[6] Schiffman, S. S., *Odors: Evaluation, Utilization, and Control*, Annals of the New York Academy of Sciences, Vol. 235, 1974, pp. 164-183.
[7] Yoshida, M., *Japanese Psychological Research*, Vol. 6, 1964, pp. 115-124, 145-154.
[8] Schiffman, S. S. and Erickson, R. P., *Physiology and Behavior*, Vol. 7, 1971, pp. 617-633.
[9] Moskowitz, H. R. and Pell, D., "Individual Differences and Congruences in Odor Perception," paper presented at the Eastern Psychological Association meeting, New York, 1975.
[10] Moskowitz, H. R., "Congruence of Aromas and Aroma Descriptors," paper presented at the meeting of the Institute of Food Technologists, Chicago, Ill., 1975.
[11] Carroll, J. D. and Chang, J. J., *Psychometrika*, Vol. 35, 1970, pp. 282-391.
[12] Coombs, C. H., *A Theory of Data*, Wiley, New York, 1964.
[13] Stevens, S. S. and Galanter, E. H., *Journal of Experimental Psychology*, Vol. 54, 1957, pp. 377-410.
[14] Stevens, S. S., *Science*, Vol. 118, 1953, p. 576.
[15] Kruskal, J. and Carmone, F., "How to Use M-D-SCAL and Other Useful Information," Bell Telephone Laboratories, Murray Hill, N.J., unpublished manuscript.
[16] Alabran, D., Moskowitz, H. R., and Mabrouk, A. F., "Carrot-root Oil Components and Their Dimensional Characterization of Aroma," *Journal of Agricultural and Food Chemistry*, 1975, in press.
[17] Moskowitz, H. R., "Individual Differences in Odor Quality Perception," unpublished manuscript, 1975.
[18] Moncrieff, R. W., *Journal of Physiology*, Vol. 125, 1954, pp. 453-465.
[19] Amoore, J. E. and Venstrom, D. in *Olfaction and Taste II*, T. Hayashi, Ed., Pergamon, London, 1966, pp. 3-18.
[20] Wish, M. and Carroll, J. D. in *Sensory Evaluation of Appearance of Materials, ASTM STP 545*, P. N. Martin and R. S. Hunter, Eds., American Society for Testing and Materials, 1973, pp. 91-108.
[21] Amoore, J. E. in *Olfaction and Taste III*, C. Pfaffmann, Ed., Rockefeller, New York, 1969, pp. 157-171.

J. J. Powers[1]

Experiences with Subjective/ Objective Correlation

REFERENCE: Powers, J. J., "Experiences with Subjective/Objective Correlation," *Correlating Sensory Objective Measurements–New Methods for Answering Old Problems, ASTM STP 594,* American Society for Testing and Materials, 1976, pp. 111–122.

ABSTRACT: Practical knowledge acquired during a decade of devising methods for the correlation of sensory evaluation for flavor with gas-liquid chromatographic analysis of food volatiles is described. The evolution of methods, precautions that should be taken with each method, limitations of the methods, and minimizing of experimental error on the sensory and objective sides are discussed.

KEY WORDS: sensory mechanisms, stepwise discriminant analysis, multivariate analysis, analysis of variance, chi-square vlues, correlation, contingency tables, headspace analysis, food flavor concentrates, potato chips, model odor systems, grits, canned blueberries, blueberry-whey beverage, gas chromatography

This paper chronicles experiences in the development of methodology. It is intended to be practical by specifically pinpointing errors, progress in overcoming limitations of methodology, and precautions which should be kept in mind with any type of subjective/objective application.

With respect to flavor or odor analysis, two things have made possible the correlation of sensory evaluation with objective measurements. One is the development of computers which, of course, applies to all areas being discussed at this symposium. The other is the development of gas-liquid chromatography because gas chromatography has made possible the gathering of a great mass of information about volatile compounds, many of which are flavor compounds. The problem is: of all the peaks observed, which ones are telltale as to quality differences. Thus, the problem turns out to be one of multivariate analysis.

When gas chromatography first originated, there were many predictions that the mysteries of food flavor would soon be unraveled. The assumption was that the flavor of a given food depended upon one or a few compounds

[1]Professor, Department of Food Science, University of Georgia, Athens, Georgia 30602.

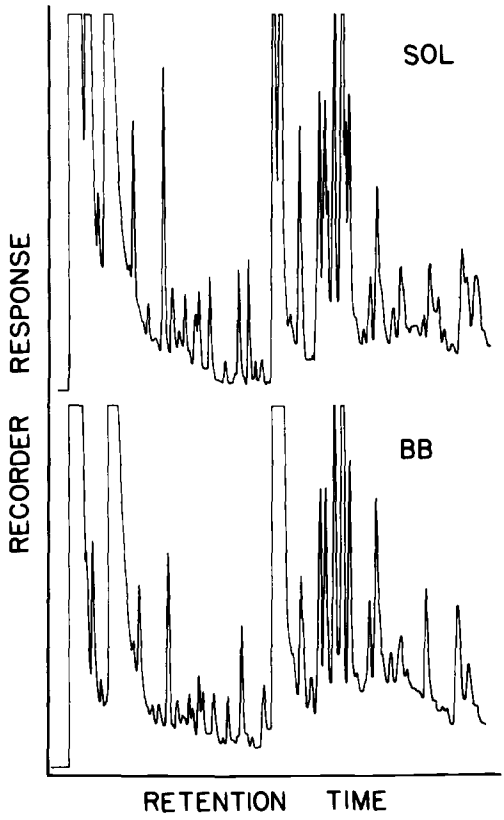

FIG. 1—*Comparison of chromatograms for extracts of Sol and BB canned blueberries.*

unique to that food. Actually, when one examines the kinds of compounds identified in many different types of foods, one is struck by the number of compounds which are common to these foods. Correlation of sensory quality with chemical composition is essentially a matter of pattern recognition, much like associating names and faces. Once in a great while a given individual will have one feature so striking that that feature alone is enough to identify him or her; but more commonly we identify individuals by putting together bits of information such as the color of one's eyes, hair, height, body configuration, and many other characteristics.

Figure 1 illustrates the same kind of a problem as applied to flavor. The two chromatograms of Fig. 1 are not identical, but one has to scrutinize them carefully to discern differences. The two chromatograms are from two varieties of blueberries, Southland (Sol) and Briteblue (BB), that differ slightly in flavor. Only by summing up the differences can one decide whether the chromatograms represent different specimens, and this cannot really be done by visual inspection.

Let us look at means of differentiating among specimens.

TABLE 1—*First "step" of a stepwise discriminant classification of blends of fresh and stale potato chips (ground), nine specimens of each blend.*

Blends		Blends					
Fresh	Stale	100% Fresh	80:20%	60:40%	40:60%	20:80%	100% Stale
100%	...	①	3	2	1	2	0
80%	20%	0	⑧	0	1	0	0
60%	40%	3	3	⓪	2	1	0
40%	60%	0	2	2	⑤	0	0
20%	80%	0	2	1	5	①	0
...	100%	0	1	0	0	0	⑧

Variable	*F*-Value
Peak 2/Peak 1	34.59

Stepwise Discriminant Analysis

In 1966, the application of stepwise discriminant analysis to correlate flavor with chemical analysis was first described [*1,2*].[2] Actually, for three or four years prior to that time, the problem had been probed in an effort to break away from the idea that a given food flavor is the result of a few characterizing substances [*2-5*].

Table 1 shows the results obtained when Miller [*2*] attempted to classify six blends of fresh and stale potato chips ground together. The reason for grinding was to have homogeneous specimens for sensory evaluation. The ratio of Peak 2/Peak 1 was the best single predictor, but as may be seen from Table 1, classification of the specimens was miserable. Table 2 shows the results after five steps, that is, the use of five variables. Classification was somewhat better but still short of being useful in a practical way; however, it should also be pointed out that the flavor differences were not so great that the panel could distinguish among the chips at each blend level. The chief flaw in the objective classifications shown in Tables 1 and 2 was the quality of the chromatograms. Rather simple headspace sampling methods, unlike the sophisticated sampling-analysis technique described earlier by Dravnicks,[3] were used.

Miller's procedure [*2*] was a considerable advance over methods published prior to that time and even more so were the results of Keith [*6*] and Powers and Keith [*1,7*]. Fully successful classification by stepwise discriminant analysis may be seen in publications from our laboratory [*7-10*].

One theory we had in the beginning was that all possible single ratios should be generated from the peaks [*1-9*]. We have since observed that this is not generally necessary, that direct use of peak areas is usually sufficient; but we have occasionally returned to the use of peak ratios when a given task could not be accomplished through the direct use of peak areas [*11*]. The

[2]The italic numbers in brackets refer to the list of references appended to this paper.
[3]See pp. 5-25.

TABLE 2–*Classification at Step 5.*

Blends		Specimens					
Fresh	Stale	100% Fresh	80:20%	60:40%	40:60%	20:80%	100% Stale
100%	...	⑦	0	2	0	0	0
80%	20%	3	⑤	1	0	0	0
60%	40%	1	1	④	0	3	0
40%	60%	1	1	1	③	3	0
20%	80%	1	0	2	0	⑥	0
...	100%	0	0	0	0	0	⑨

Variable	*F*-Value
Peak 4/Peak 1	4.48
Peak 3/Peak 2	1.21
Peak 1/Peak 3	1.73
Peak 4/Peak 2	1.15
Peak 4/Peak 3	0.60

ratio is the area of one peak divided by that of another. One advantage high speed computing affords is that ratios and many other kinds of transformations can be generated for trial to seek out the most suitable means of classification or prediction.

One final word with regard to stepwise discriminant analysis. Preferably, one should do a step-down analysis instead of a step-up analysis as we did. In theory, the first two variables taken because each of them has a high *F*-value does not necessarily mean these two variables, as a pair, are the best predictors. There may be some other pair which is actually better though the two components of the pair are not significant variables separately.

Stepwise analysis is highly useful if one wishes to classify specimens. Its chief disadvantage is that if one wishes to predict the identity of a new specimen, one has to run the analysis all over again merely to add one more specimen to the classification.

Multivariate Analysis

Multivariate discriminant analysis has some advantages stepwise analysis does not have; although in terms of a step-down analysis, there are many things in common. The procedure we now follow is to generate a discriminant function such as

$$Z = \lambda_1 X_1 + \lambda_2 X_2 \ldots \lambda_k X_k$$

where

Z = weighted mean,
X = area of each peak, and
λ = weighting value.

From the computer program MUDAID [*12*], output such as is illustrated

TABLE 3—*Weights and correlations of discriminant function.*

	Peak Areas									
	PK05	PK07	PK09	PK10	PK18	PK19	PK22	PK23	PK26	PK31
Weight	0.459	0.715	-9.83	-17.7	-18.4	4.96	-0.868	18.7	24.9	-35.5
Correlations	0.42	-0.12	-0.12	-0.08	-0.21	0.24	-0.07	0.15	0.38	-0.28

in Table 3 is obtained. Table 3 is actually for data already partially edited, that is, some of the peaks detected have been dropped from consideration because they were not correlated with product differences. This is done by running a one-way analysis of variance, and all peak areas not exhibitng a significant F-value across the products are dropped. Peak areas that do differ significantly with product differences are then used to calculate a discriminant function such as shown in Table 3. The correlation coefficients are examined next, and then the function is edited by eliminating those peak areas least correlated with the discriminant function. For example, Peaks 7, 9, 10, 22, and 23 might be eliminated. If so, then a new discriminant function has to be calculated because the weighting values change some, according to the particular variables used. Upon editing, the discriminant function in Table 3 became

$$Z = 0.47X_5 - 21.6X_{18} + 4.98X_{19} + 25.5X_{26} - 37.3X_{31}$$

From the discriminant function, the weighted means (Z) for each product are calculated. Table 4 shows calculations for four odor mixtures which were examined organoleptically. To evaluate new specimens, the peak areas of the new (or unknown) specimen are substituted in the equation for the X's, the weighting values previously calculated are used; and from the resulting Z value, an estimate is made of the quality (or the identity) of the specimen. Table 5 shows results obtained when 20 unknowns were identified from gas-liquid chromatography patterns by statistical analysis. One specimen that was actually Specimen R was predicted to be Specimen B but even this is within the 0.05 level of probability, that is, 1 in 20 specimens. The predictions in Table 5 are based on single chromatograms; had the predictions been based on replicate chromatograms, the one specimen that was misclassified certainly would not have been so badly misclassified, and it might

TABLE 4—*Relation of weighted means of the discriminant function to the odor scores of four model mixtures.*

	Mixtures			
	R	G	P	B
Odor score	2.32	2.70	2.93	2.90
Weighted mean	2.75	2.48	2.27	2.17

TABLE 5—*Success in identifying 20 odor mixtures based on gas-liquid chromatography and multivariate analysis.*

Mixture	*Z* Values Calculated From Known Specimens	*Z* Values of Specimens
R	2.75	2.73, 2.67, 2.75, 2.14, 2.74
G	2.48	2.49, 2.49, 2.59, 2.49, 2.48
P	2.27	2.27, 2.27, 2.25, 2.25, 2.26
B	2.17	2.19, 2.16, 2.14, 2.14, 2.16

not have been misclassified at all. Success in predicting unknown food specimens can be seen in prior publications [*13,14*].

There are some other things that can be done to add assurance to the predictions made, aside from increasing replication. One is to calculate more than one equation. In ranges where the difference is not great (see the *Z* values of Mixtures P and B, Table 5, for example), a second discriminant equation may be calculated for that range only [*14*], or, alternately, one may calculate an entirely different equation using some transformation of the original variables. Table 6 illustrates this procedure. Both procedures have value. Calculating a second, or even a third, a fourth, or a fifth equation is no chore. The usual calculation takes from 7 to 80 s of computer time.

Objective Errors

In discussing headspace analysis of the potato chip specimens, reference was made to poor resolution of specimen differences. This is the kind of error with which all gas chromatographers are familiar, and all endeavor to detect as many peaks as possible and to resolve them fully. There is another type of error which should be kept in mind. Figure 2 shows chromatograms for the model odor system referred to above. The Chromatogram RM is for the liquid Mixture R; RV-0.5 is a chromatogram of the vapor in the headspace of a vial held at 0.5°C. Chromatograms RV-22 and RV-72 are for headspace samples drawn from vials at 22 and 72°C, respectively. Note that no peaks are detected in the liquid sample prior to about 14 min, but even at 0.5°C peaks become evident at 1 to 4 min retention time; and they are clearly evident for headspace samples drawn at 22 and 72°C. The importance of this relates to correlation with sensory analysis. The things that enabled the panelist to distinguish among the samples and thus to assign preference scores may not have been the nine main compounds used to make the mixture, but traces of impurities in the headspace. Because of greater volatility, they are enriched in the headspace whereas they were present in too small an amount to appear on the chromatogram for the liquid sample.

One thing that needs to be stressed is that one is not evaluating a food or odor specimen as it is perceived by the panelist. There is no method of extraction or analysis which yields chromatograms that are mirror images of the ratios of volatiles in the food or odor substance itself. Like funny mirrors

TABLE 6—*Alternate Z values calculated to aid in resolution of specimen differences.*

	Mixtures			
	R	G	P	B
Z values calculated from % area of nine peaks	2.75	2.48	2.27	2.17
Percent differences	10.9	9.2	4.6	
Z values calculated as the arc-sine transformation for five peaks	0.57	0.32	0.09	.01
Percent differences	78	356	900	

which make people ludicrously fat or skinny through distortion of their image, any chemical method of analysis distorts the ratios of volatiles as they exist in a food itself. Very useful correlations may be obtained, but one should never forget that the relation between things perceived sensorially and measured objectively is arbitrary at best.

Sensory Errors

Except with reference to the potato chips above, the discussion so far has been as if all error were on the objective side. As a matter of fact, sensory

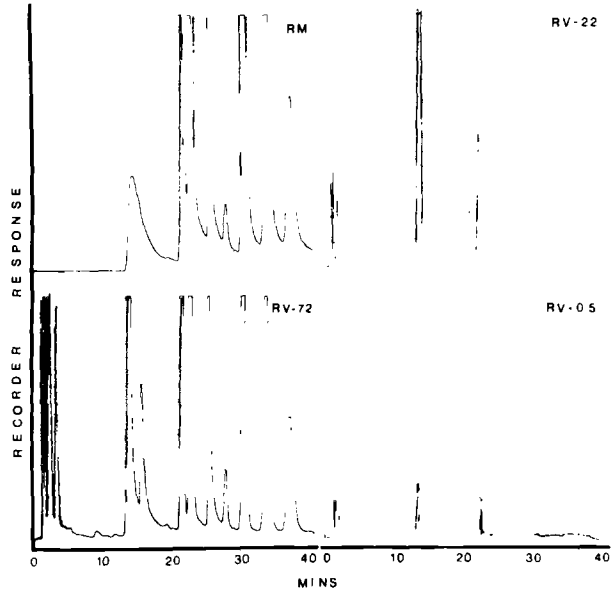

FIG. 2—*Comparison of chromatograms for a liquid sample, RM, and headspace vapors over the liquid sample when the temperature of the liquid was 0.5, 22, and 72°C, respectively.*

TABLE 7–*Correlations among judges evaluating the flavor of grits.*

Judge	Judges								
	2	3	4	5	6	7	8	9	10
1	-0.42	-0.45	-0.42	0.22	-0.25	-0.23	-0.46	0.63	-0.39
2		0.08	0.39	-0.66	-0.23	0.50	0.47	-0.27	0.32
3			0.38	-0.12	0.27	-0.06	0.42	-0.61	0.53
4				-0.30	0.27	0.15	0.42	-0.64	0.57
5					0.29	0.64	-0.21	0.11	-0.32
6						-0.25	-0.24	-0.21	0.12
7							0.10	-0.01	0.05
8								-0.67	0.57
9									-0.61

NOTE–Subset A = Judges 1, 5, 9,
Subset B = Judges 2, 4, 7, 8, 10, and
Subset C = Judges 3 and 6

error is usually much greater. If one is determining differences among specimens, there are means of assessing the consistency and the discrimination power of the judges. The task is a little more complicated when preference or acceptability is involved. Additional considerations then come into play. If an individual prefers something, no one can say he is wrong. Preference is a personal matter. One thing we do is to calculate the correlations between each judge and all other judges [*11,15*].[4] Table 7 shows the correlations for 10 panelists who evaluated four varieties of grits produced in different ways. There were 35 panelists all together, but the correlations are listed only for the first 10 judges to keep the table within reasonable bounds.

If one examines Table 7, one sees that the preferences of Judge 1 are negatively correlated with those of Judges 2, 3, 4, 6, 7, 8, and 10. In a like manner, Judge 2 is negatively correlated with Judges 5, 6, and 9. If one examines only the panel means and whether or not the means differ significantly, one may be mislead as to preference for the product. Table 8 shows the consequences of failing to look behind panel means. If one were to use only the panel means, one would conclude that blueberry-whey beverage made with 33 to 40 percent blueberry pulp is preferred, but actually a sizeable portion of the panel, 38 percent, preferred the beverage with only 19 to 26 percent pulp. As far as the eight judges are concerned, no one is more

TABLE 8–*Relation of compatible subsets of judges preference for blueberry-whey beverages.*

Beverages, % Pulp Level	Preference Scores		
	Subsets A (5 Judges)	Subsets B (3 Judges)	Panel Means
19	2.3	2.7	2.5
26	2.6	2.8	2.7
33	3.6	1.9	3.0
40	3.7	1.6	2.9

[4]See pp. 56–72.

TABLE 9—*Partition of judges into subsets having compatible preferences for the acceptability of grits.*

Subset	Percent of Original Panel
A (11 judges)	31.4
B (4 judges)	11.4
C (2 judges)	5.7
D (2 judges)	5.7
E (1 judge)	2.8
F (15 judges)	42.8
	99.8

"right" than other. All were significantly good judges except they happen not to like the same thing.

Partitioning such as is done in Table 8 is often useful to explain why correlations between gas-liquid chromatography and sensory data are not high. Differences in flavor preference may account for poor correlation.

Table 9 shows how 35 judges partitioned out when overall acceptability of grits was evaluated. The partitioning is perhaps excessive because the panelists were separated so as to be completely compatible in their preferences within each group. If one were to tolerate some incompatibility in preference, the number of subsets could be reduced. The differences in overall acceptability come about because panelists consciously or subconsciously may not consider flavor, mouthfeel, color, and appearance to be of equal importance.

Aside from explaining laboratory results, partitioning of the judges into compatible sets enables one to get an idea of the percentage of the population which would prefer or at least find acceptable a given formulation. Of the 35 judges, 15 (subset F) had no preference. As far as these judges are concerned either the differences between specimens are not great enough or else they are nonsensitive tasters. It would make no difference to these judges which formulation was used. No one would make marketing decisions on the basis of 35 judges, but correlating the judges into comparable sets does enable one to estimate the percentage of the population a given formulation might please. Not only should the judges in the majority as to preferences be counted but also those who have no preference because presumably they would not object to the product.

Interval Between Scores

A second problem that has plagued those carrying on sensory evaluations is the fact that panels may not maintain constant intervals between score levels. One of the benefits we have been able to derive from a new computer program (PREPRO), which was originally set up to maximize the distant between treatments, is to find out whether or not the panels are maintaining equal intervals between score levels [*16*]. As may be seen, when the values are

TABLE 10—*Order and distance between score levels, grits, 16 judges.*

Original Score Levels				
1	2	3	4	5
Transformed Acceptability Scores				
-0.935	-0.752	0.014	1.48	1.63
	0.183	0.766	1.466	0.15
Transformed Flavor Scores				
-1.48	-0.585	-0.783	0.672	2.04
Transformed Mouthfeel Scores				
-1.04	-0.953	-0.476	0.493	1.89
	0.087	0.477	0.969	1.40

transformed so as to normalize the data, the intervals between scores are no longer the same. One should thus apply least-significance-difference or Duncan's multiple-range tests with caution because one has already learned that the distance between scores has a different meaning to the panelists, depending upon the part of the scoring range under consideration. Alternately, the transformed score should be used to make any calculations. This is what we regularly do. Note the inversion for two levels for flavor in Table 10. These two levels should be combined, that is, the score level compressed to four levels because the panelists are not differentiating between scores 2 and 3.

Contingency Tables

The last thing is more a matter of a visual aid than something which cannot be attained by other means. There are tables to estimate whether or not differences between specimens are significant, but one advantage of the PREPRO program is that one has this information displayed as a part of the output of the contingency tables between treatments and the various scoring levels. What the program does is to construct a contingency table between the score values actually assigned and the treatments. The chi-square value for the distribution is then calculated and from that one obtains a printout as to the probability of this distribution being a random one. As may be seen on the left side of Table 10, the probability that scores would have been assigned in the manner they were on a random basis on only 1.96×10^{-12}. On the right side of the table a random distribution is shown. The computer program first constructs a contingency table based on the actual assignment of scores, then it randomizes the data, holding the sums for the columns constant and

keeping the sums for the rows as constant as possible. The two probability statements can tell whether the panel is approaching randomness in its assignment of scores or the differences between treatments are so great that it is highly unlikely that the assignment of scores is a random process (see Table 11).

In conclusion, our experience is that one can be quite successful in correlating sensory and objective measurements. Chromatograms for foods known to differ in sensory qualities may be disconcertingly alike, but by multivariate statistical means one can generally winnow out those factors that correlate with sensory quality. Critical examination needs to be applied to the sensory data too, for it is the cause of some of the difficulty commonly encountered in correlating sensory and objective measurements.

The question may arise as to whether we correlate the sensory scores directly with the chemical-physical measurements. We do not. Both types of measurement are dependent variables, that is, the properties of the products determine the sensory and objective values. The two types of variables are correlated (see Table 4) but only indirectly. In some instances, both sensory and objective values may be used in the discriminant equation to effect maximum resolution among specimens. An example would be the use of an appearance score where no corresponding objective test exists. If one is seeking to minimize the use of sensory panels because of their cost or the difficulty of assembling suitable panelists, and provided suitable objective tests exist, then naturally the equation will generally be based on purely objective values. Multivariate statistical analysis usually permits one to select a combination most suitable for the task at hand.

TABLE 11—*Comparison of panel assignment of acceptability scores for grits and random assignment, holding column sums constant.*

Panel Assignment[a]							Random Assignment[b]						
Products	Scores						Products	Scores					
A	2	4	15	29	14	64	A	13	24	14	14	7	72
B	9	25	13	13	4	64	B	9	20	15	5	5	54
C	16	25	14	5	0	64	C	13	17	15	15	4	64
D	10	22	9	17	6	64	D	15	19	17	16	5	72
E	20	26	14	3	1	64	E	12	24	7	16	9	68
F	17	19	14	6	8	64	F	12	21	11	7	3	54
	74	125	79	73	33	384		74	125	79	73	33	384

[a] $r = 0.484$,
$r^2 = 0.235$
chi-square ($DF = 20$) = 101.93, and
significance level = 1.96×10^{-12}.

[b] $r = 0.157$,
$r^2 = 0.025$,
chi-square ($DF = 20$) = 17.49, and
significance level = 0.621.

Acknowledgments

Appreciation is expressed to the Frito-Lay Division, Pepsico Company, Inc., for yearly support of a graduate assistant, which has made possible sustained research on subjective-objective methodology.

References

[*1*] Powers, J. J. and Keith, E. S., "Evaluation of the Flavour of Food from Gas Chromatographic Data," Abstracts of 2nd International Congress of Food Science and Technology, Warsaw, Poland, Aug. 1966, p. 439.

[*2*] Miller, W. O., "Application of Gas-Liquid Chromatography to Determination of Potato Chip Staleness," Ph.D. thesis, University of Georgia Library, Athens, Ga., 1966.

[*3*] Dawes, J. C., "Gas Chromatographic Analysis of the Volatile Components of Canned Red Pimientos," thesis, University of Georgia Library, Athens, Ga., 1964.

[*4*] Wannamaker, W. G., "A Study of the Relationship between Gas Chromatographic Analysis of Pimentos and Organoleptic Preference Tests," thesis, University of Georgia Library, Athens, Ga., 1964.

[*5*] Stephens, C. E., "A Study of the Relationship between Gas Chromatographic Analysis of Canned Peaches and Organoleptic Preference Tests," thesis, University of Georgia Library, Athens, Ga., 1967.

[*6*] Keith, E. S., "Investigation of Chemical-Mathematical Methods of Evaluating Food Flavor," thesis, University of Georgia Library, Athens, Ga., 1966.

[*7*] Powers, J. J. and Keith, E. S., *Journal of Food Science*, Vol. 33, Feb. 1968, pp. 207-213.

[*8*] Powers, J. J., *Food Technology*, Vol. 22, April 1968, pp. 39-44.

[*9*] Young, L. L., Bargmann, R. E., and Powers, J. J., *Journal of Food Science*, Vol. 35, 1970, pp. 219-223.

[*10*] Powers, J. J., *Proceedings*, Third International Congress of Food Science and Technology, Washington, D.C., Aug. 1970, Institute of Food Technologists, Chicago Publishers, pp. 395-402.

[*11*] Powers, J. J. and Quinlan, M. C., *Journal of Agricultural and Food Chemistry*, Vol. 22, Sept./Oct. 1974, pp. 744-749.

[*12*] Applebaum, M. and Bargmann, R. E., "A Fortran II Program for MUDAID, Multivariate, Univariate and Discriminant Analysis of Irregular Data," Technical Report NONR 1834, No. 39, 1967, University of Illinois, Urbana, Ill.

[*13*] Milutinović, L., Bargmann, R. E., Chang, E-Y, Chastain, M., and Powers, J. J., *Journal of Food Science*, Vol. 35, 1970, pp. 224-228.

[*14*] Powers, J. J., Cox, R. J., Eiland, M. C., and Ackroyd-Kelly, P., *The Flavour Industry*, Vol. 2, Feb. 1971, pp. 87-92.

[*15*] Quinlan, M. C., Bargmann, R. E., El-Galalli, Y. M., and Powers, J. J., *Journal of Food Science*, Vol. 39, 1974, pp. 749-799.

[*16*] Kundert, D. R. and Bargmann, R. E., "Tools of Analysis for Pattern Recognition," Themis Report No. 22, 1972, Department of Statistics and Computer Science, University of Georgia, Athens, Ga.

Summary

Each of the papers is both a review of the state of the art and an exposition of methods. A primary purpose of the symposium was not merely to point up developments in multivariate analysis and correlation methods but rather to evaluate the strong and weak points of each method and to explain how one actually applies the methods. Thus, considerable detail is given to the rationale for each step or alternate procedure to aid those not familiar with a given field in understanding the purpose for each operation. This is especially true for the first paper because it was designed to be a case study of a typical problem.

The problem itself involved trying to develop objective tests which could be used to predict the effect of compositional or manufacturing changes on the sensory quality of beer. Odor dimensions evaluated included intensity, pleasantness/unpleasantness, and different degrees of intensity for each character note discernible. In product development, it is important to know the effect that changes in raw materials or manufacturing process will have upon desired sensory qualities. This may be done by experimentation, aided by intuition and elementary statistical analysis, but a better way is to document the properties of the components of the product objectively so that one can use the vast resources of chemistry to predict what probably will happen if the chemical composition is changed in a certain manner. If the data are not analyzed fully, one may miss useful correlations or arrive at hasty decisions. Because the odor of most foods consists of hundreds of compounds, multivariate analysis is needed.

For the type of problem described, stepwise regression analysis was most suitable. Regression analysis is desirable when variables are continuous. If attributes were being evaluated which fall into distinct classes, then stepwise discriminant analysis would be a more suitable choice. For either type of analysis, an attempt is made to classify all the samples available into different groups on the basis of the variable which appears to distinguish among the different groups most effectively. If this variable is insufficient to permit the various specimens to be classified, then a second variable is entered into the equation and an attempt is made to separate the specimens anew. This process is carried on until (*a*) the specimens can be grouped into distinct sensory classes or (*b*) further analysis would be fruitless because the remaining variables would not add further to the resolving power of the equation.

Analysis, of course, does not start with curve fitting. A computer program to organize the gas liquid chromatographic data in logical groups according to area size, retention times, frequency of peaks within a given retention-time range, odor descriptions attached to each peak, and other pertinent information is first used. Likewise, computer analysis is used to learn whether the variables assume some known mathematical function such as being linear, logarithmic, or parabolic. There are standard programs to perform the various tasks mentioned previously and to carry on the regression analysis. All one has to do is punch out a few cards to instruct the computer as to what to do.

Once a formula is generated, it should be tested against a new set of specimens to be sure that the equation has general properties and is not merely provincial for the set of specimens from which it was originally derived.

Techniques of headspace analysis were described as were "sniffing" procedures. Information can often be obtained about the compound(s) causing a particular peak more simply by sniffing than by any other method.

Through multivariate analysis to ascertain the correlation coefficients between the various substances and sensory qualities, more straightforward procedures can be followed to modify a food so as to change its flavor in a desired direction than can be done by pure sensory experimentation.

In the first paper and the others, statistical design was discussed. The second paper was devoted primarily to designs for sensory analysis. Paired comparisons, triangle tests, multiple paired comparisons, and other designs as well as various scoring or rating systems were discussed and illustrated. For some purposes, highly trained judges like tea tasters or wine tasters are used. For many purposes, laboratory-size panels are employed. These judges receive training too and their responses have to be evaluated to learn if they are sufficiently sensitive and consistent, but the training period is shorter than for the expert type where many years of apprenticeship may be involved. Consumer panels are sometimes used. The panelists are not usually instructed, other than explanation of what is being asked of them, because one wants their response to be typical of consumers, not modified by training. For the laboratory panel, proper facilities must be provided so that extraneous noise, odors, or other factors will not distract the panelists from their task.

A paper on sensory-objective analysis of canned blueberries applied some of the concepts described above. Panelists were asked to rate blueberries for acceptability, flavor, mouthfeel, color, and appearance and to answer certain questions relative to statements about the products. The panelists were examined for reliability and sensitivity after the investigation. Their responses were subjected to analysis of variance and only those whose F values for acceptability were at the 0.05 level or higher or whose F values for flavor, mouthfeel, color, and appearance were at the 0.01 level were retained as judges. In other words, if they did not meet the criteria just mentioned, their responses were discarded. The study on canned blueberries paralleled in some respects the study on beer. It was pointed up that there has been a gradual

change from esoteric manipulative models to exploratory techniques such as correlations between best discriminators and observable variables, finally to regression and simple covariance models. In the blueberry study, regression analysis was not used because there was no way to relate sample to sample or replicate to replicate since no one can of berries sufficed for all the sensory and objective tests. Contingency analysis and analysis of variance was used to relate qualitative and quantitative ratings.

The methods available to carry on sensory analyses are fairly well standardized, though, of course, refinements in methods continue to be made. The kinds of objective tests which should be applied to a given problem are more difficult to categorize since rarely does any one objective test encompass the same scope as does the corresponding sensory phase. For example, mouthfeel results from many elements such as viscosity of the food, force required to compress it or shear it, particle size, and hardness. No one instrument performs all these functions. Thus, considerable thought has to be given to physical or mechanical measurements which hopefully parallel, to a certain degree at least, ordinary chewing or mouth movements and force. The selection of suitable objective tests was covered by a speaker who has long experience in the design of various mechanical devices and in comparisons of their function with corresponding sensory analysis. Shear force measurements, compression force, and other objective tests were amply explored. Experimental design was likewise discussed, especially limitations which must be considered.

One paper was more fully oriented to design problems than the others. This paper dealt with response surface methodology as applied to textile fabrics. It was divided into three parts: (1) planning stage, the designing of an experiment which will produce the desired results; (2) methods which can be used to provide a numerical measurement of the subjective response; and (3) example in textile research to illustrate the principles and techniques. The second part mentioned previously dealt with the devising of some numerical scaling procedure for factors which themselves are not numerical, such as the luster of a fabric. Formulation of an equation does not just happen; careful planning is required to ensure that the data will allow a formula to be generated. The prospective levels of dosage or treatment must be chosen with care so that the data points will permit the true shape of the curve to be estimated precisely. Also discussed and illustrated was the need to restrict variables or replications so as not to wear out judges or tie up production lines too long when regular production must be suspended to prepare the experimental specimens. Response surface experiments require thoughtful planning, but if properly performed, a graphical display which can be derived from the data provides good guidance in selecting variable settings to produce the desired response.

A paper on univariate functions in psychophysical measurements traced the history of the development of metric approaches to the correlation of sensory responses with chemical or physical measurements. Category scaling and mag-

nitude estimation were compared. The value of being able to predict consumer response through application of psychophysical laws and the economic benefits to be derived through an understanding of the rate of response to concentration changes for different sweeteners were used to illustrate practical applications of mathematical modeling.

The discussion which generated the greatest controversy was multidimensional scaling. Some of the statisticians present felt that internal analysis or principal component analysis rests upon shaky foundations. On the other hand, some who have used multidimensional scaling consider that the method has value in permitting an investigator to see at a glance geometric relations among concepts or components. An example with odorants and linguistic descriptors was used to illustrate multidimensional scaling. Rather strong views about the suitability of multidimensional scaling in the food or odor fields has likewise occurred at other meetings. This is an area where statisticians and those with an immediate problem to solve need to get together to determine if there is not some middle ground which is sound both in theory and in application.

One paper was purely pragmatic in nature. It dealt with the developments of methods, false starts made, and the benefits derived from sensory-objective correlations.

The final paper reviewed multivariate methods from a systemic point of view. The various techniques were discussed and simple examples were used to establish the rudimentary concepts. Factor analysis, cluster analysis, and multidimensional scaling were compared. Applications were generally drawn from the tobacco industry. Of particular value is a glossary of technical terms and a list of available computer programs.

Index

P

T