J. Soc. Cosmetic Chemists, 16, 763–776 (1965)

Principles of Consumer Product Testing

JEAN F. CAUL, Ph.D., and SHIRLEY A. RAYMOND, B.S.*

Presented May 4, 1965, New York City

Synopsis—After a review of the basic features of consumer panel testing, the requirements for a discriminating panel are described. The utility of this panel, especially in combination with a "Use Profile," is demonstrated with several examples. It appears to be entirely possible to relate consumer findings to laboratory measurements. Such data can guide reformulation or may be the go-ahead signal for marketing of the product.

Introduction

The chemists engaged in the development of new products and the statisticians responsible for the evaluation of these creations for the consumer market recognize the many difficulties in designing the perfect mathematical model for successful forecasting. These obstacles include considerations of cost and speed and of such concepts as statistical probability, quota sampling, Rorschach test, computer programming, decision-making, concept testing, and share of market. Another difficulty arises from the difference in emphasis between those trained in the physical sciences and those trained in mathematics, social psychology, and business. The chemist frequently emphasizes the product's properties; the mathematician stresses the exact and deducible relationships between quantities and operations; the social psychologist probes the explanations of group behavior; and the business-trained person is concerned with who exchanges money for merchandise.

* Arthur D. Little, Inc., Food & Flavor Section, Life Sciences Division, Cambridge, Mass. 02140.

Another difficulty which limits the correlation between product properties and consumer preference originates with marketing management. It is market-oriented rather than research-oriented and hence views the successful achievement of this correlation of various disciplines as virtually impossible.

Nevertheless, this laboratory's experience in the testing of product models and final formulations suggests the possibility of relating measurable product properties to consumer acceptance and rejection and thus sustains the hope that a product's specifications can be so drawn up in the laboratory as to make virtually certain its acceptance by consumers. The purpose of this paper, then, is to outline the use of consumer panels as product evaluation instruments. The first part of the discussion is concerned with the various kinds of consumer tests and discusses the do's and don'ts of consumer product testing; the last part explains methodology and the rationale underlying the suggested approach.

Tests Involving Consumers

Market Tests

Basically there are only two types of consumer tests: One in the marketplace where consumers exchange money for merchandise; the other, where they do not. Marketplace tests obtain information about the performance of the packaged, labeled, and priced product in actual sales situations, and such tests fall properly within the scope of sales development and market research people.

In all other consumer tests, that is, where consumers do not hand over their own money for products, information is obtained in a test situation about the consumers or the product or both. A test situation does not equate with a sales situation, but it eliminates some of the guesswork and, therefore, some of the risks in launching and merchandising a new product.

Consumer Testing

When information is obtained only about consumers, the approach should be called consumer testing. Like market testing, consumer testing or the testing of people falls within the scope of persons involved with sales and frequently needs to rely on the knowledge and training of social psychologists as well. It may involve hypothetical circumstances, as in evaluating a product concept when no product exists.

It may involve tricking consumers by having them evaluate the same product in three differently branded containers. It may aim to characterize consumers through depth interviews. Whatever the approach, consumers are under scrutiny, not the product.

Consumer Product Testing

Consumer product testing, on the other hand, should aim to test the product. At present there is some confusion of objectives about product testing because of two factions: the marketing people and the technical people. Marketing people have the task of taking the product, packaging it attractively, naming it, selecting a desirable label, getting it into distribution channels, developing enticing advertising copy, and pricing the finished product—all this by a target date that probably was set as soon as the green light was given to the technical staff to develop the product. Little wonder that they want to modify the product test with marketing-type questions. But all too often the marketing people appear not to remember that their best efforts will come to naught if the product is not right. They can attract first purchases, but they cannot bring about repeat buying unless the product they are trying to sell is in fact acceptable to consumers.

On the other side, the technical people have the responsibility of creating the product, and they necessarily have a different set of questions to be answered: Does the product meet its concept or design conceived for it? If not, how far along is it? What are its negative features? What are its positives? If it does meet the concept, is the concept acceptable? Is the product right?

In the authors' opinion, a consumer product test should test the product, and only if this primary objective will not be sacrificed may marketing-type questions be included in the test.

Components of Consumer Product Testing

In planning a consumer product test, one must consider four test components: 1) size and characteristics of the consumer panel; 2) number of different samples to be tested; 3) length of use of the product; and 4) method of obtaining information from the consumers. The key to selecting the best choice for each component is this: What must be learned about the product? In other words, this test is designed around the problem.

The Panel

Ordinarily, the selection of a panel of consumers is based on the premise that a consumer panel should be representative of the total population, and, therefore, statistical analysis can be applied to the findings. From such analyses, then, it should be possible to ascertain the probability or reliability of the findings.

There are two kinds of population samples to which statistics are justifiably applied: Random samples, that is, population samples selected without a definite pattern; and quota samples, in which the people who make up the test panel fit the characteristics of a predetermined pattern—such as proportionate representation of the various income and educational levels of a standard metropolitan area (SMA) as determined by the U. S. Bureau of the Census. Inasmuch as only persons who are willing to participate in a test are represented, no sampling is truly representative. Nevertheless, statistical analysis can be and is used successfully in making decisions by persons who take into consideration this and various other gaps in their methodology. Too often, however, statistical analysis is misapplied or interpreted incorrectly.

The larger the sample the greater the probability that the findings are not due to chance. But, as the sample size increases, so do the costs of the test. Therefore, depending on available funds and the degree of risk one is willing to take, a particular sample size is selected.

In addition to size and general representative qualities of the panel, specific characteristics of a panel could be deliberately selected according to the product's end use. For example, a hair conditioner for beauty shop use would be tested by a panel of beauty shop operators on their clients. Suppose that three popular hair conditioners are already on the market for beauty shop use. One could adjust the composition of the testing panel so that users of these first, second, and third place rank products would be proportionately represented. This subsampling technique would yield more information on the performance of the hair conditioner in comparison to its ultimate competition.

Different Samples to be Tested

Often technical people wish to obtain guidance from small consumer panels regarding the direction to be taken with an unfinished product. If this single product were given to the consumers, they would exaggerate its unfinished qualities and fail to see or evaluate its positive

features. Consumers, like people, emphasize the negatives. One solution is to give the consumer panel two products, both unfinished but having different negatives and positives.

Whenever two products are submitted to consumers, they will form the parameters of a comparison test. If one is a good product and the other only a fair product, the comparison will tend to emphasize the extremes; in this instance, the consumers would regard the good product as excellent and the fair one as poor. In any paired comparison both products should be of the same general quality.

It would be difficult to test two shampoos with the same consumer panel. An acceptable solution to this problem is to use two similar panels, one for each product. The panels should be alike in all pertinent variables: the same numbers of users and nonusers of cream rinses, for example; the same age distribution patterns; the same numbers of home permanent users; the same numbers of nonusers; and the like. When the returns are in, the balance of the two panels should be confirmed, and their findings can be compared.

Multiple product testing with the same panel can be done only if sufficient time elapses between the use of each product. The time lapse prevents carryover from one product to the next. Second, the time lapse must not exceed the user's ability to remember and compare the various products. Multiple product testing is one of the most misused techniques in food product testing for several reasons: the test designers seem to feel that by alternating the position of use of each product with succeeding consumers, they have canceled out product interactions and carryovers. Such is not the case. They also seem to feel that consumers remember each product's properties equally well. This is not the case either. Consumers do a paired comparison of the first two products, sort out the major differences, and then assess the succeeding products according to the major differences they remembered from the first two. The test designers frequently make another mistake by forcing consumers to respond to three or more products tasted at one session, headcounting these preference rankings and misapplying statistical analyses, coming up with probability figures that are accepted as true. The point here is that statistics should not be applied to data from improperly conducted tests. Fortunately, most cosmetic products require repetitive use before they can be truly evaluated by consumers, and multiple cosmetic product testing is therefore contraindicated.

With random panels of consumers who run the gamut in intelligence, there is always some danger of introducing unwanted biases induced by sample codes. A panel of intelligent consumers obviates the need to be deeply concerned about findings based on a like or dislike of the code used for an otherwise unidentified sample. But from common sense codes should be avoided that might take on meaning, e.g., the letter X, the letter-number code A-1, the butter-score numbers 88 or 93 for margarine, or G-11 for a soap. Most test designers avoid color codes and simple letter or number codes; they usually use two- or three-digit number-letter combinations.

Use Period

In food testing if only a simple "yes/no," "go/no-go" kind of result is needed, a product may be subjected to "one-shot" testing. This type of test is usually conducted in trafficked areas, such as a store, county fair, or bus terminal. This test resembles the man-in-the-street opinions poll. Persons who happen to pass by and who have the time and inclination are invited to participate. Often they are asked to choose between two samples on a preference basis. Considerable numbers of consumers can be reached by such a test method, but it should be remembered that the data relate only to first impressions.

Conceivably the aromas of two perfumes could be checked out in this way, if the bias inherent in such a consumer test population were recognized. But the actual properties of the perfumes and the consumer's changes in attitude could be learned only through a repetitive use test. For most products there is a first impression, a get-acquainted period, and a final impression. The final impression leads to final assessment, which in the marketplace would determine second purchase.

A use test, then, should be conducted for a long enough period to attain a final impression. This might be as short as three days for an after-shave lotion or as long as four weeks for a scalp conditioner. It is difficult to sustain a panelist's interest for more than two weeks unless re-stimulation is offered. Recently the authors were able to hold a home-use consumer panel of 200 persons together for three months by check postals, fresh samples, and friendly letters. The check postals really served a dual purpose. They activated the panelists and provided information on changes in attitude and frequency of product use so that one could be reasonably certain of obtaining final impressions at the conclusion of the test.

Methods of Obtaining Data

There are only two basic methods for obtaining responses from consumer panel members. The first is by questionnaire and the second by face-to-face or telephone interview. Even the interviews are commonly conducted by questionnaire. Usually these are structured questionnaires, whereby all interviewers use the same words, the same questions, the same procedure.

Interviews, understandably, are more expensive than questionnaires mailed to and from the consumer panel members. Interviews provide several advantages. First, they assure a response, whereas a mailed-in questionnaire can be totally ignored. Second, they assure early responses. A mailed-in questionnaire can be put aside until the respondent has forgotten details about the test product or until after the deadline for compiling and reporting results has passed. Third, more questions can be asked by an interviewer. Less patient consumers will not spend the time to fill out carefully a long and complicated questionnaire; a carelessly answered questionnaire is worthless.

In using the interview technique, considerable effort should be made to obtain reliable, honest interviewers and then to train them so as not to allow them to introduce a bias. In general, most testing agencies do not identify the client or the underlying reasons for testing the product. This eliminates another source of interviewer bias. A part of the interviewer training period is devoted to pre-testing the questionnaire, and if the interviewers are intelligent people, they may be able to point out ambiguous words, better sequences of questions, and suggest where structured probes could be inserted.

Questionnaires to be filled in by the consumer panel members also should be pre-tested. With experience the composer of questionnaires learns how to ask questions so that any literate person will understand them. But the danger of questionnaires lies in not knowing the meaning of terms, either those asked by the questionnaire or those used by the consumers. For instance, checklists in questionnaires have asked for preference on mildness. Does this refer to feeling sensations, or to flavor, or to level of flavoring? When a consumer says a toothpaste leaves his mouth clean, is he referring to cleansing ability or refreshment due to the flavoring? When he says there is too much carbonation in a beverage, does he mean just that or is he saying that the weak flavoring is subjugated by the carbonation?

Not knowing the meaning of terms results from not knowing the

product's properties. People who really know the product should participate in developing the questionnaire and in editing the responses before compilations are made. In fact, a great deal can be learned by reading through all answers from each respondent. This provides a perspective of the over-all issues before the specifics are examined, as well as an understanding of what each panelist is trying to say.

Requisites of Pilot Consumer Product Testing

The foregoing discussion of the component parts of consumer product testing was intended to orient the many facets of product testing from a technical person's point of view. The topography of consumer product testing is very similar to that of conducting a laboratory analysis. First, define the objective: What information is needed and with what precision? Second, select the measuring device and analytical procedure: What precautions and controls will be needed? Third, conduct the analysis and obtain the findings. Fourth, relate the findings to the objective.

This approach to product testing is always used in A.D.L.'s Food and Flavor Laboratory. During these studies several important principles have been developed which experience has shown are the most useful for product development problems. In the following discussion, it will become evident that the primary principle in applying consumer panels to product testing is to understand what is to be tested and why.

The Discriminating-Communicating Panel

Usually a random or quota sample is not used. Because of the desire to have accurate information that can, if necessary, be translated into technical terms, a panel of consumers is selected who have these particular qualifications: First, they are interested in testing the particular product. This means that they will probably complete the test, taking care to respond carefully to the questions and giving the product a fair trial. While aware that they are performing a special favor, the testers also feel they are influencing the design of products that they, as consumers, may someday see in the marketplace. Second, this consumer panel is intelligent. This means that they will follow our instructions, try to avoid confusing product identities or codes, and likely will be able to express themselves adequately. Their ability to communicate their observations is of paramount importance. Third, they observe or distinguish a product's properties accurately. This ability to discriminate has been demonstrated by pre-testing.

Then, as mentioned earlier, panelists will be selected according to specific characteristics. In testing a new instant coffee, care was taken to include families who used only instant coffee, families who used only brewed coffee, and those who used both. The "instant-only" users showed they had become accommodated to the flavor of the then insipid but not unpleasant instant coffees, while persons who were familiar with brewed coffee recognized the virtues of the flavor characteristics of the new product. This illustrates selection of panelists according to type of product they use.

Frequency of product use can also be a criterion of selection. In another test it was found that frequent users of the current product were strongly against a variant of the product, while occasional users were delighted with the variant. This could mean two products with an over-all increase in product use.

More attention is being paid to teen-age products, and the authors have a source of panelists, who have been found to be as discriminating as their parents and often more communicative.

Size of panels may be as small as 25 persons or families and as large as 100 persons or families. Since the size of the A.D.L. panels is not large, these tests are called pilot tests. The roster from which panelists are usually selected consists of the families of A.D.L. Cambridge employees who have expressed their desire to participate in product tests. Most of them have lived in the area more than five years. Naturally, then, the first question to face is: Is this a regional panel? The answer generally is "No."

If the type of product is used nationally—e.g., mouthwash—then its use properties can be evaluated by a discriminating-communicating panel anywhere. If the product is designed for a specific region in the U. S., this panel can indicate if the product has the qualities it is supposed to have. They may not particularly like the product, but they can isolate its elements.

The second question to face is: How do the results obtained from discriminating-communicating panels compare with those from a national and not necessarily totally discriminating panel? To probe this question, a paired comparison of two toothpastes was put through the A.D.L. consumer toothpaste panel and a national panel twice as large. The preference trends and reasons for preference were the same from both panels, but in analyzing the questionnaires, the information from the A.D.L. panel was found to be more definitive. This result was not unexpected.

But there is an even more important rationale to the use of discriminating panels. Because they do discriminate, they provide the basis for a rigorous test. If the product is acceptable to persons who can discriminate, its properties should also be acceptable to persons who cannot or do not discriminate.

Prerequisites

To plan a proper test—that is, to select the consumer panel, to decide how often the product should be used and for how long, to choose the method of obtaining responses from the panel, to develop instructions for the consumers, and to anticipate the terminology they might use—the designers of these tests charge themselves with two responsibilities: to be sure of the purpose of the test and to be sure of the test product's properties.

Every test is specifically designed around the product and the test objective. If the product is a model of a concept, the product development group may wish to know if it matches the concept and, if not, what modifications are needed. If it is to be a new product, is it in its present status acceptable as a whole; does it have more positive than negative features? If a variant of a currently marketed product-type, how do its attributes compare with those of the marketed product; is there a positive that could be exploited in advertising? If it is an improved version of an existing brand, do consumers see the difference and do they consider it an improvement; do they see it as a major or minor improvement?

Once the test objective is defined, extensive effort will be made to define the product. First, if such technical information is not already available, technical analyses or examinations of the product will be carried out. Since the authors' consumer product testing mainly concerns foods and other flavorful products, this technical examination, while including observations on pH, color, and viscosity, will mainly be a Flavor Profile. Flavor Profiles are produced by experienced panels, who work under controlled conditions and use standardized techniques for smelling and testing. The Flavor Profile is a tabular record of the product's sensory (aroma and flavor) properties.

Having completed the Flavor Profile, the panel members will work to produce a use profile. For a soft drink, they would drink (as opposed to taste) a bottleful of the beverage in much the same way that consumers will—for example, gulp it down, pour it over ice, let it warm up in the glass, and drink it from the bottle. The value of a use profile is

incalculable. It bridges the gap between the technical analysis and the consumer responses, so that the test operators will be able to relate consumers' descriptive terms to the use profile and finally to the technical or flavor analysis. To report consumer findings to the product developer, one must be able to speak in his technical terms.

In addition to use profiles, the panel will also produce abuse profiles. Such information enables one to characterize the inherent latitude of the product and also to anticipate untoward responses and to guard against them. If, for example, the directions for preparing a soup call for 10 minutes' simmering, the effects of under- and over-simmering should be known. Undercooking could cause the noodles to be tough and the flavor underdeveloped. Overcooking could concentrate the soup, making it strong and salty.

The Testing Situation

During this study period, the product elements that consumers are likely to observe are sorted out, and decisions are then made regarding the test method. As previously mentioned, if the product is not a finished one but is to be tested in order to obtain guidance for its future direction, one might decide to test it in comparison to another unfinished product. If it has a counterpart on the market, then its performance could be tested in comparison to the "blind" marketed product. And similarly, if it has been designed as an improved product, the assessment of its improvement features could be made through a comparison with the unchanged product. A new-concept product would of course be tested by itself; in fact, single-product testing may also be applied to any product if the test can be designed to meet its objectives.

The product's properties and its intended use will determine some of the instructions for the consumer panel as well as duration of the use period and, therefore, the supplies needed. A side-by-side comparison would be requested in a statistically designed difference test aimed at defining flavor attributes of puddings differing only in sweetener content. This technique provides for a minimum time lapse and thus a direct comparison, which is a stringent test. Stronger tasting products, such as mouthwashes, would not be amenable to such immediate comparisons. Instructions would therefore request alternate use of one mouthwash on one day and the other on the next day, both to be used according to the panelists' normal patterns. Alternating-day use gives a closer comparison than alternating-week use, for example.

Obtaining Responses

The test objective, the product's use properties, and its intended use will determine how to obtain information from the consumer panel. If, as in foods, eating quality is the primary product property to be tested, responses can be obtained by questionnaire. If, as in hard liquors, other unknown attributes may supersede flavor, face-to-face interviews are in order. Interviews are worth their higher costs for the advantages already cited, i.e., they are time saving and provide more definitive information.

The interviewers are personable technical people, who are made to be fully cognizant of the purpose of the test, the client, the product's Flavor Profile, and its use and abuse profiles. These technical people are experienced flavorists, members of the A.D.L. Food & Flavor staff who train in on the particular test. During their training period, they use the product as the consumers will and, under supervision by their peers, practice first among themselves and then with consumers. In their practice sessions they learn to establish rapport, how to conduct an unstructured interview on this product, and to train their memories so as to be able to write up the interview or fill in their data sheets away from the scene of the interview. In other words, they learn to have a conversation about the product, allowing the consumer to describe her own impressions without channeling her responses by checklist type questions. They rely mainly on open-end questions and are allowed to pursue whatever the consumer considers important, probing on the spot for clarity and definition of descriptive and vague terms.

Similar precepts guide the development of questionnaires to be filled in by the consumer panelists. The note transmitting the questionnaire tells the consumer in general terms why the test has been conducted. After asking about the use of the product, the questionnaire itself asks mainly open-end questions about the product. If the panelists have been stimulated to be communicative, they will give their appraisal, sort out the favorable features and those that, in their opinion, need to be improved, and indicate the relative importance of the features they have discussed. This result, however, is not left to chance. A pretest of the instructions, use period, and questionnaire is almost invariably conducted, using about five families.

Other things not left to chance concern the test samples themselves. Before they are placed for home use, they are sampled at random and checked out to make certain that they do represent the product to be

tested and that they do not vary significantly. After they have been placed, other samples are periodically checked in the laboratory so that any unanticipated changes occurring during the testing period will be known.

Interpretation of Consumer Responses

Since in most of this work the response sheets are a series of short essays, each one is read for meaning and perspective. The respondent, whether interviewer or consumer, assumes this will be done. For example, for Soup A the consumer may have written that its unfavorable feature was "not salty enough," and for Soup B, "good rich flavor." Since the Flavor Profiles showed that both soups had the same salt level, she is not saying she would like to taste salt. She is stating that Soup A's flavor needed something and to improve it she would have added salt. If her descriptions for each soup had been read separately, Soup A might have been tallied under "salt level low"; this would be inaccurate and misleading.

The value of use profiles for interpretation is illustrated by the following example. In a series of paired comparisons involving six balanced panels, the objective was to select the best flavored product of three. Each product had been tested against the other and also against a control. The six sets of response sheets were read separately, and it was found that Product B evoked a seemingly different reaction when tested against Product A than against the control. Against Product A it was called flat; against the control it was called pleasant tasting. Without the use profiles of the products, one could not have interpreted these findings. But they dovetailed nicely with the use profiles of A *vs.* B, A *vs.* control, and B *vs.* control, which showed each pairing had different flavor parameters. Product A's over-all flavor was stronger, more identifiable but less appropriate; the control's flavor had several negative components; Product B's flavor was weaker than A's, stronger than the control's, and more appropriate than either. In other words, the test situation was different with each pairing, and the use profiles had defined the differences.

A research study of an oral product is a final example of relating consumer findings to product properties. This study was aimed at defining the product's important flavor elements and the ranges wherein these elements could make positive and negative contributions to preference. First, the consumer panelists isolated the important elements. Then, in a series of paired comparison tests conducted over a two-year period, the

intensity of each flavor element was varied separately, each time relating the consumer findings back to the test products' Flavor Profiles. After only six tests the study furnished the flavoring formulators with a Profile-blueprint. This tabulation defined the flavor character notes, their upper and lower intensity limits, and their order of sequence. The formulator with his expert knowledge of flavoring materials could then create new models which he was able to evaluate at the bench.

Thus, it is possible to relate consumer findings to laboratory measurements in order to draw up a product's specifications in meaningful technical terms. If the basic principles of good testing are followed when discriminating consumers make up the test panel, a product that fails to pass a consumer product test can be purposefully modified; and a product that passes is ready to be turned over to the marketing people.