Paired Preference Tests

In consumer research, the role of the paired preference test appears to be sacrosanct, whether deserved or not. Paired preference simply means putting one product up against another, and instructing the respondent to indicate which one the respondent prefers. The 'preference' measure does not show how much one product is liked (or how much more one product is liked versus another). Rather, the paired preference test is simply a head to head match, with 'winner take all' (on at least a person by person basis). The results of paired preference tests are reported in percents, rather than in degree of liking (as would be reported in a scaling exercise).

Paired preferences tests can extend to other attributes because the researcher can also ask the respondent to indicate which product has more of a specific characteristic. The characteristic need not even be a sensory one (such as depth of color, thickness, graininess of texture, etc.). Rather, the characteristic could even be an image one (e.g., one product is 'more masculine').

Paired preference tests, popular as they are, provide relatively little information. The conventional (yet unproven) wisdom is that consumers make choices by implicitly comparing products to each other. Thus, the typical paired test pits a new product against either the gold standard that it is to replace, or against the market leader (and only the market leader) against which it is thought to compete.

The positives of paired preference testing are the simplicity of the test in the field execution, and the ease of understanding (viz., a paired comparison result). The negatives are that the demand for head-to-head comparison may focus attention onto small, irrelevant differences that exist (especially when the respondent is searching for a hook on which to hang the comparison), and that the data cannot be used for much beyond the comparison results. Paired comparison data are not particularly useful for product development because they give no real guidance.

It is worthwhile digressing for a moment here to understand a little of the intellectual history of paired testing, because of the widespread use of the methods, and the limitations (which do not appear to affect the use or misuse of the procedure). Paired testing got its start more than a century ago. The German psychologist, physiologist, and philosopher, Gustav Theodor Fechner (Boring 1929), was interested in the measurement of sensory perception. However, according to Fechner, the human being is not able to act as a simple measuring instrument in the way that we understand these instruments to operate. [Today's researchers, especially those in experimental psychology and psychophysics, would vehemently disagree with Fechner, but keep in mind that we're dealing with the start of subjective measurement, not with its well developed world today]. According to Fechner, one way to measure sensory perception was to ask people to discriminate between samples. From their behavior Fechner was able to determine the magnitude of difference needed between two samples to generate a difference. The psychometrician, L.L. Thurstone (1927) carried this analysis one step further by developing paired comparison methods, in which respondents judged which sample was stronger, heavier, liked more, etc. From the paired comparison data (the ancestor of our paired preference tests), Thurstone developed a subjective scale of magnitude. Thurstone's scaling methods were soon adopted by researchers to erect scales of sensory magnitude and liking, from which developed the long-standing acceptance of paired methods in applied product testing. Thurstone, publishing in the academic literature, was thus able to influence subsequent generations of applied market researchers, many of whom do not know the intellectual origins of this test.

Sensory Questions – How Much Of A Characteristic Does My Product Possess?

Respondents can act as judges of the amount of a characteristic. Sometimes these characteristics or attributes can be quite simple – e.g., the sweetness of a beverage. Other times the attributes may be more complex, such as the ‘tomato flavor’, which calls into play a host of sensory attributes. Still, other times, the attributes may be simple, but require explanation (e.g., the ginger burn of a ginger candy).

There is an ongoing debate in the scientific literature (especially fragrance, but also food) about the degree to which a respondent can validly rate sensory characteristics. On one side of this dispute are those who believe that the only attribute that a consumer can validly evaluate is liking. These individuals believe that it is improper to have consumers rate sensory attributes. On the other side of the dispute, are those (including this author) who believe that a well instructed respondent can validly rate sensory attributes, and indeed such a respondent (not an expert, mind you) can switch focus from liking to sensory to sensory directional (see below), for many attributes.

The dispute is simple enough to solve through experimentation. In cases where experts and consumers rate the same products it has been shown that their ratings correlate with each other (Moskowitz 1995), and with known physical variations of the product (Moskowitz and Krieger 1998). The scientific literature suggests that consumers can validly rate sensory attributes, and that their attribute ratings line up with known physical changes in the stimulus. Literally thousands of scientific articles in all manner of disciplines related to sensory perception suggest that the unpracticed respondent can assign numbers whose magnitude matches the physical magnitude of the stimulus (see Stevens 1975). One could ask for no clearer validation of a respondent’s abilities.

Sensory Directionals or "Just Right" Information – What To Do

Quite often in developmental research a key question is ‘what’s wrong with this product (if anything), and what does the consumer feel that we should do to correct this problem’. When a respondent is asked to evaluate problems, the typical question is known as a sensory directional. The question may be phrased somewhat as follows: "Please describe this product: 1=far too dry… 5 = Perfect on dryness/wetness … 9=far too wet". Note that the respondent is assumed to know both the ‘ideal’ level of dryness/wetness, and the degree to which the product deviates from that ideal.

Respondents often find this type of question fun to answer, because the question allows them to become experts by telling R&D product developers what to do. Sensory directionals are surprisingly ‘on target’ for visual attributes (viz., the respondent does know the optimal or ideal level of darkness or stripes), usually on target for texture, but sometimes on target, other times off target for taste/flavor. The directions are generally off target for certain types of emotion-laden attributes such as ‘real chocolate flavor’, perhaps because these attributes are hedonic attributes (liking), disguised as sensory attributes. No chocolate ever has enough ‘real chocolate flavor’. Even if the developer were to use a lot of chocolate flavoring, the product would still taste too bitter.

Typically, product developers use ‘rules of thumb’ by which they translate these just right scales to direction. Thus when a respondent says that a beverage lacks ‘real fruit flavor’, the product developer knows that often the respondent means that the product is not sweet enough – or, that changing the amount of sugar will change the fruit flavor.

Where Do Sensory and Other Product Test Attributes Come From?

Since much of product testing involves finding a way to have consumers communicate with product developers, it is important that the questionnaire cover the key sensory attributes of the product. A glance at different questionnaires will reveal a remarkable range of attributes, from the very general (overall appearance, aroma, taste, flavor, texture), down to the very specific (e.g., amount of pepper flavor, even amount of black pepper flavor). Some questionnaires are filled with specifics; other questionnaires appear to be probing the surface, without much depth in the quality of information being generated. Table 2 shows an example of an attribute list for beer (Clapperton, Dagliesh and Meilgaard 1975).

Scales – What's The Right Metric To Measure Subjective Responses?

Over the years researchers have fought with each other, often passionately, about the appropriate scales to use. We saw above that there are a variety of scales to measure liking. We are talking here of a more profound difference in scaling – the nature of the scale, and the allowable transformations. The standard nine point hedonic scale (Peryam and Pilgrim 1957) is an example of a category scale. Category scales are assumed to be interval scales – the differences between adjacent category points is assumed to be equal up and down the scale. However, there is no fixed zero point. Interval scales (like Fahrenheit or centigrade) allow the researcher to do many statistical analyses, such as calculating the mean, the standard deviation, perform T tests, regression, etc. The interval scale has no fixed zero, however, so that one cannot calculate ratios of scale values. On a nine-point scale of liking we cannot say that the liking of 8 is twice as much liking as the liking of 4. Other scales, such as ratio scales (with a fixed zero) do allow the researcher to make these ratio claims. Weaker scales, such as ordinal scales (viz., rank order these products in degree of liking), just show an order of merit. One cannot even talk about the differences being equal between ranks 1 and 2 versus ranks 2 and 3.

In the end, most researchers end up with the scale that they find easiest to use. The category scale is by far the most widely used because it is simple and can be anchored at both ends. [Sometimes the researcher anchors every point in the scale as well]. Many academic researchers use ratio scales to study the magnitude of perception, and from time to time ratio scales have been used in the commercial realm of product testing. Many researchers in business often use rank order or ordinal scaling. They have to be careful not to interpret differences in ranks as reflecting the magnitude of differences in subjective magnitude.

Product to Concept Fit

Beyond simply measuring acceptance or sensory attributes, researchers want to discover whether or not a product and a concept fit together. The concept sets the expectations for the product. When a consumer buys a product there are often some expectations about the appearance, the taste, the texture of the product, along with non-sensory expectations (e.g., aspirational expectations, such as sophisticated, etc.). Products can be tested within the framework set up by these expectations in the product-concept test.

The actual test execution is quite simple. In one variation the respondent reads the concepts; an opportunity to form expectations is given. The respondent then evaluates one or several products. For each product the respondent rates the degree to which the product delivers what the concept promises. The scale is also simple – either the product delivers too little, just right, or too much of what the concept promises. The scale sounds easy enough to use, and in actuality it is easy (at least respondent say that they have no problems).

One of the key issues to keep emerging in the evaluation of product/concept fit is whether or not respondents really have an idea of what a product should taste like, smell like, perform like. If the concept promises a specific flavor (e.g., an Italian flavor), or a specific performance (e.g., 366 MHz processor in a computer), then respondents can easily ascertain whether or not the product lives up to the concept. For many aspirational concepts, however, such as those developed for perfumes, it is difficult, if not impossible, to truly show that a product and concept agree with each other. The respondent may try to articulate the reasons for the concept/product fit, but the reasons will be different for each person.

Base Size – How Many Respondents Are Enough?

The issue of base size continues to appear in product testing, perhaps because base size more than any other factor influences the cost of a project. The greater the number of respondents in a study (viz., the larger the base size), the more comfortable the researcher should feel about the results – presuming, of course, that the respondents participating are the correct respondents. This has led to recommendations or at least rules of thumb dictating 100+ respondents for conventional product tests, but far more (e.g., 300+) for claims substantiation tests (Smithies and Buchanan 1990).

Base size issues are not as simple as they might appear. The statistician would aver that the reason for the large base size is that it reduces the uncertainty of the mean. The standard error of the mean, a measure of expected variation of the mean, were the study to be repeated, drops down with the square root of the number of respondents. That is, one feels more comfortable about obtaining the same mean the next time if one uses a large enough base size in the first place. Matters can get out of hand, of course, if this requirement for uncertainty reduction in itself demands a base size so high that the test is unaffordable. [Many novice researchers, in fact, often become so fearful about the uncertainty that they refuse to do studies unless there is an enormous base size, preferring rather to do qualitative research instead].

Another way to look at base size considers the stability of the mean rating with an increasing number of respondents. The first rating is, of course, the most influential rating in determining the average, for it is the only data point. Each additional rating affects the average less and less (viz. by 1/n, where n=number of ratings). The author has shown that the mean stabilizes at about 50 ratings, whether the attribute deals with liking, or with sensory judgments, and that this number, 50, appears to hold even when the data comprises ratings from a homogeneous population of individuals with the same preference pattern (Moskowitz 1997). Therefore, it appears that the researcher will not be in particular danger if the base size exceeds 50. A base size of 100 might be better (or at least appear so psychologically), but the mean will not change much from 50 ratings to 100 ratings, assuming the sampling rules are maintained for choosing the respondents.

Part 3 – Package Design Research

A Short Overview

Package research is often the neglected child in consumer testing of price, positioning, product, and package. Until recently, little attention had been given to the product package. Companies might launch new products, and involve their in-house package design professionals as well as outside design houses. However, much of the research done was either qualitative (focus groups) or none at all. Part of this lack of research stems from the perception that the package is not particularly important – it is just a vehicle by which to enclose the product (especially in fast moving consumer goods). Part of the lack of research can be traced to the design houses, which perceive their work as artistic, incapable of judgments. [No design house wants consumers to design the package for them]. As a consequence, there is a paucity of research literature on package design, although there are scattered articles here and there, and a number of well respected companies specializing in package research. [We can compare this dearth of literature to the extremely high volume of literature on advertising testing, perhaps because many more dollars are spent on advertising, so it is immediately more important to be ‘right’].

Today, however, with increased competition in many categories, the package is assuming renewed importance, and researchers are rising to the occasion. Manufacturers are sponsoring short in-house courses on the importance of package design. Package design firms, once the bastion of the artist, are welcoming quantitative research. [It should be acknowledged that these design firms always welcomed qualitative research, because that research, like qualitative in advertising, probed and revealed aspects of the package that were important for the designer]. Package design firms are also adopting a more holistic approach to the design process. According to researchers at Cheskin (a brand identity and design firm), commenting on the design of a beer package (1999), "Beyond helping you manage clients, research can actually help you create better design – not by dictating color and form, but by informing your intuition…When your job is to make a product sell itself at the point of sale, how do you know that your design will deliver the goods?...What does the package say about the beer inside?…How should your client position the product?…If you just look at the packages, what would you say about the beer?"…

What Should Package Testing Provide?

Package testing serves a number of purposes. The most important purpose for package testing is to confirm or disconfirm the objectives of the package designer. Typically package designers create packages (either graphics and/or structures) with some objective in mind – such as reinforcing the brand, communicating new benefits, enhancing the chances that the product will be selected. In all of these objectives package testing must provide some idea as to whether or not the new package is successful. A modicum of sensitivity to the creative process is also in order, since package testing often reflects on the creative abilities of the designer – who is both an artist and a business-responsive individual.

Focus Groups

For many years, the conventional package evaluation consisted of focus groups. Focus groups, properly conducted, can be very valuable in the creative process, especially in the up-front developmental stage. The focus group is private, does not come out with hard and fast results, and can cover a great deal of territory in a warm, informal manner. In a package design focus group the respondent can verbalize reactions to package features, can identify key features that elicit interest, and can talk about the coherence between the package itself and the brand. A sensitive package designer gets a great deal out of focus groups because the designer can see and hear how the consumer reacts to the packages. By presenting different possible packages, the designer can see which ones ‘work’ and which do not. Focus groups provide the designer with a great deal of feedback, generally in a non-threatening, non-judgmental manner (viz., not judged by another professional, although consumers could reject the package). It should come as no wonder that this type of qualitative research has been welcomed by designers, because in essence it reflects how they would intuitively go about obtaining feedback about their creations.

Focus groups can, however, backfire in several ways. First, they can be misused. In a focus group people say many different things, and the listener can select specific phrases to suit his/her own purpose and agenda. Second, and even more importantly, respondents often like to ‘play designer’. Respondents like to tell designers what to do – even if the designers aren’t really interested in following the sage advice. (Glass 1999). Consequently, there may develop an antagonism between client/respondent (with the client wanting to follow the respondent's suggestions) and the designer (who has a basic artistic and business idea in mind and simply wants feedback).

Profiling Packages

At the simplest level the researcher can determine whether or not the package is acceptable to the consumer, or fits the brand. This type of testing typically involves attitudinal measures. The researcher shows the package to the respondent, and obtains a profile of ratings, similar to the way that the researcher obtains product ratings. The key differences are that the researcher may obtain profiles of the 'expectation' of the product (based upon the package), as well as ratings of the package itself. The attributes can vary substantially from one product category to another. Some of the ratings may deal with interest in the package (or in the product, based upon exposure to the package). Other ratings may deal with one's expectation of the product based upon the package. If the respondent actually uses the product in the package, then the researcher can obtain ratings of person-package interaction (ranging from ease of carrying or gripping the product; ease of opening; ease of removing the product; ease of storing the package, etc.). This type of information is extremely valuable to the package designer, who wants to find out whether or not the package is 'on target'.

Typically the designer receives only the simplest of "'briefs' or project description, such as the requirement that the package be upscale, that it live up to the brand, and that it communicate an effective or good tasting product. The data from the evaluative package tests provide the diagnostics to demonstrate whether or not the package as designed actually lives up to the requirements set by the package design group at the client.

Behavioral Measures – T-Scope and Speed of Recognition

At a more behavioral level the package testing may involve behavioral measures. One behavioral measure is the ability to identify whether or not a package is actually on the shelf, during a very short exposure time. The rationale behind this type of testing (called T-Scope or tachistoscope testing) is that the typical shopper spends relatively little time inspecting a store shelf. In order for a package to make its impact it is important that the package 'jump off' the shelf (visually). The researcher assumes that those packages that are perceived in the short interval permitted by the T-Scope have shelf presence. In some demonstrations, Elliot Young (1999) has demonstrated that in these T-scope tasks the speed at which the stimulus information is available is often so low that the respondent has to use color and design cues, rather than brand names. Well-known brands with distinctive package features (e.g., Tide® brand detergent, with its concentric halos) are rapidly recognized, even if the brand name is incorrect. If the research interest focuses on the recognizability of the single package, then one can present single packages and determine the fastest shutter speed (viz. the least time needed) for the package to be correctly identified. If the research interest focuses on the 'findability' of the package on the shelf, then the researcher can place the test package at different locations on the shelf and then parametrically assess the contribution of package design and location on the shelf as joint contributors to "findability".

Eye Tracking and The Features of A Package

Eye tracking is another method for testing the shelf. The typical shelf is an extremely complex array of package design over which the consumer's eye wanders. The objective of package designers is to guide the consumer to the package. This action then constitutes the first step in selecting the product. Eye tracking allows the researcher to identify the pattern that the follows. Is the eye drawn to the client's particular package? Does the eye wander away from the package (thus diminishing the chances that the customer will select the product)? Young 1999 has also demonstrated the use of eye tracking technology, based upon military research developments in the 1970's. The eye tracking technology traces the location of the line of site. It shows how the consumer, when presented with either a shelf set or a single package, explores the stimulus, tracks how the eye wanders, and records what the eye sees. When done for a single stimulus (e.g., an over the counter medicine), the eye tracking technology can show if and when, and for how long key messages are looked at (but not whether these are good or not). When done for the entire shelf the eye tracking technology can identify whether or not the package is even looked at, and for how long. Based upon these analyses, Young and his associates have amassed a variety of generalizations about where on the shelf the package should be, how many facings the package should have, etc.

Optimizing Package Design – Can Art and Science Mix?

One of the most intriguing developments is the advance of conjoint measurement (systematic stimulus variation) into package design. Conjoint measurement comprises the experimental variation of components in a concept in order to understand the underlying dynamics of how the components perform. The respondent evaluates full combinations of these components (e.g., benefits, prices, etc.), and from the ratings researchers estimate the part-worth contribution of each component. The same research paradigm (viz., systematic variation of components) has begun to enter package research, but this time the components are features of the package (e.g., different names, colors, graphics, etc.), and the concepts are full packages comprising these systematically varied features. From the responses of consumers to test packages (e.g., created on the computer screen), the researcher and package designer quickly discover what every design feature contributes to consumer interest, communication, etc.

Some of the power of the conjoint approach applied to graphics design can be seen by inspecting Figure 4 (showing template, components, one finished package, and a table of utilities). The conjoint system enables the designer to rapidly understand the effectiveness of each concept element, and then to create new and better combinations, by incorporating high performing graphic elements, and by discarding poor performing elements. Of course attention must always be paid to the artistic coherence of the design. Unlike concept testing using conjoint measurement, however, there is far less in the literature (and in actual practice) using systematically varied package features. The best explanation for this is that package design is still an artistic endeavor, resisting research on the one hand, and yet welcoming research insights on the other.