Product Testing by Mail in a Panel[1]

Mainly we are testing consumer goods such as food, beverages and cigarettes; personal care and hygiene products; household cleansers and care products. We have also carried out tests of packages, of concepts, and publications; consumer usage & habit studies; image surveys.

The general experiences have been:

1. *Great care is required in wording of questions and design of questionnaires, so that they are easy to understand and to be answered.*

2. *Almost all kinds of questions can be used in mail surveys.*

   Exceptions:
   Questions requiring spontaneous reactions cannot be used, since people tend to think before filling in a questionnaire.

   Questions regarding knowledge: People will try to inform themselves on the subject so as not to appear stupid.

   Questions which influence each other. Either respondents look through the whole questionnaire before answering it, or they discover that a later question deals with a subject already touched, and they may then correct their first answer.

3. *All kinds of scales that are being employed in interviewer-surveys can be used if set up and explained properly.*

4. *If questionnaires are designed properly, respondents fill in questionnaires very properly. They are trying at least as much as normal interviewers to do a good job.*

5. *Return rates are high: In product tests they are in the 80-90 % range, in other studies between 70 and 80 % (if answering the questions requires more effort we send an incentive with it).*

Following is a description of procedures which we have developed in 25 years of operating a panel. It does not imply that these are the only or best ways of operation. But in any case of a future co-operation across national borders, it seems useful to have some basis for common ventures - and for improvements.

1.) Establishing a Panel

1.1.) Recruitment

In the very beginning the largest proportion of our households was recruited by personal interviews. We have had a national staff of appr. 800 interviewers. And we had them contact households both on the basis of random route as well as quota selection.

Generally, the contact was made in the form of a product test: The interviewer presented the product and asked whether the „housewife“ would try it and then answer a few questions. If she agreed, a few basic questions were asked, the product and a relatively short questionnaire left behind. She was to answer the questionnaire after having tried the product, and then mail it to the institute (in an envelope with postage paid for).

Usually, this was not a „real“ product test but one where we had bought inexpensive but attractive products. The concept was: First, to convince housewives that they wouldn’t be co-operating for nothing but would get attractive products. Second, to show them that it doesn’t require much thinking and effort to fill in a questionnaire.

[1] Part of the information brochure drafted 1990, and actualized 1998 for our European partner institutes

In the meantime, we use various procedures to gain new households:

- On one hand, it is quite possible to recruit additional households by asking present participants whether they could suggest other families who might be interested. This is conducted about every 5 years quite successfully.

- Our panel in East Germany has been recruited in 1990 by newspaper advertisements. This proved to be a rather successful method since the to-be-tested products „Made in West Germany“ were, of course, very attractive (more than 175.000 postcards were sent to the institute!).

- Rather continuously, we are conducting tests with baby care products. For this purpose, addresses of households with babies, aged 0-24 months, are bought (available from some professional suppliers of addresses in bundles of 5.000). They receive a first letter, explaining the purpose of such tests and asking for participation. If the answer is „yes“, they are included in the next tests.

- Last but not least, there are several hundreds of people who are asking year by year for panel membership on their own.

In our experience the rate of recruitment is between a quarter and a third of people contacted. Losses, on the other hand, are insignificant. In this kind of panel operation households may be asked to try a product perhaps 3-5 times a year, and get 3-5 worthwhile products in return; consequently one does not have the attrition problem of continuously reporting panels. Our attrition rate is less than 3% p.a.

1.2.) Obtaining Basic Data

Most product tests are to carried out in target groups, - not in samples of „everybody“. These target groups are quite often of a special character, and possibly quite small, e.g.:

- Users/consumers of a particular product („category users“), such as coffee drinkers, cigarette smokers, users of facial moisturizers, hair spray or deodorant users etc.;

- users/consumers of a specific segment of such products, such as decaffeinated coffee, light cigarettes, anti-dandruff shampoos, etc.;

- users of one particular brand, or of a group of brands within a product category;

- and/or demographic categories, such as women/men, age groups, income groups; or subgroups like women between 16 and 30.

One of the greatest advantages of a panel is that such data can be gathered once a year, stored in the Panel Data File, and used for the pre-selection of target groups whenever required.

For this, it is necessary to collect such data from each household as early as possible. In many cases the lengthy „Basic Data Questionnaire“ was included as a part of the very first contact interview plus the trial product test. Other households got it after they had answered the first

product test questionnaire. Others, after indicating their interest in participation in a telephone contact with the institute.

Unfortunately such data have to a certain degree what can be called a due-time: Buying, usage, and consumption habits in many product categories do mostly change only quite a bit; but most of all, purchase/use of **brands** changes quickly and extensively. As a consequence, this background information has to be updated once a year or so.

This is an expensive and time-consuming job. Some clients pay for their required data; other ones prefer to have target group selection costs included in their individual test costs. Some clients expect the institute to have such data available as part of its service free of charge (which means costs have to be earned as part of the overhead costs).

The importance and scope of those basic data surveys in the panels call for careful planning and execution.

1.2.1.) Which Basic Data?

It is, of course, difficult both for (potential) clients and the research institute to anticipate 1 year ahead which target group(s) data might be of interest.

On the one hand there are some clients with a limited range of products and who are rather frequently carrying out tests; it is easy to know in advance which data are likely to be required.

Most clients have a larger product range and are not testing them on a regular basis. Nobody knows for sure which kinds of tests might come up during the next 12 months. Consequently, including or excluding any target group data has to be based on intelligent guesses.

At the other extreme we have new or potential clients, where little if anything is known regarding relevant, potentially useful target group characteristics.

In the course of the last 20 years, we have developed the following policy:

1. If nothing at all is known, include at least the purchase/ consumption of the product category.

2. If space permits, ask for brands used using open-ended questions. The answers need not be coded and put into the Panel Data File but left in the questionnaires. If the need arises, questionnaires can be looked through for appropriate households or individuals.

3. Only where it is worthwhile, brands should be ascertained and stored in the Panel Data File. The use of a scanning machine and an appropriate program to read tick marks makes it possible, nowadays, to collect a greater number of brands and varieties with the aid of given lists in higher speed and at lower costs.

1.2.2.) Ascertaining Brands

Usual product tests and other surveys which are directed to specific target groups of brand users are based on questions like the following, asked by interviewers in the typical face-to-face situation

„Which brand are you using actually/usually/mainly?“

Respondents saying „yes, I do use ... (e.g. NIVEA)“ are used to be taken as **users**. The other ones are non-users. This always seemed to be a clear and doubtless definition.

Relatively unknown is that already this information rather often is *wrong*. Respondents tend to name more frequently the rather big and important brands, particularly when such an „open ended“ question is the basis for the screening of users. And if „pantry checks“ are done, often it becomes visible that *other* brands are rather or additionally present in the households.

The next problem is that after a rather short time (depending on the repurchase frequency in the particular product category) the brand might be changed - and if one asks later on a second time, one will get another answer. But: what does „usership of a specific brand“ mean in such a context?

Panel researchers are aware of this problem: People switch between brands, and although the single person will use actually another brand from time to time, the sum of all users at a given date provides rather stable market shares over time.

Ad hoc researchers like most of us are, learned to know this phenomena in the end of the seventies, when the prognosis test „Assessor“ was published (Silk & Urban, 1978):

- Users do not buy and use only one brand, but chose their actually used one out of a „relevant set“ of brands;
- the decision which one will be bought next is influenced by preferences for the different brands within this relevant set (and actual in-store advertisements, price activities & promotions etc.);
- the prognosis which brand will be bought next by one individual is difficult, due to this probabilistic situation;
- but the sum of all preferences leads to rather stable purchasing probabilities respectively buyers' shares.

That means, if one tries to set up a Panel Data File including information about „brand usage“ one must be aware of this phenomenon.

If one asks in the screening questionnaire: „Which brand are you using actually?“ - one gets an information that doubtless will be wrong within a short time and to a certain degree. People do change „their“ brands within their relevant set of brands, and if one asks in a test afterwards: „which brand are you using now“ - one will receive another answer.

The solution can only be to ask for brands which consumers consider buying in a more general way. This is the basis for our question *„which brands did you use within the past 6 months ?“*

For subsequent sample selections this means:

- The definition of purchasing within the past 6 months (p6m) is the best definition to gain something like the „relevant set“ of brands. Though switches between the brands of the individual relevant set are possible, the shares of the brands are stable over time. This makes the „relevant set“-definition favourable over main-brand-concepts or others.

- If the selection of a 50 : 50 users/non-users split is requested, this must separate those who are really non-users from such with a certain affinity towards the brand in question
  (a 50 : 50 split between users of **one** brand vs. users of **a second** brand causes more problems, and must be discussed separately).

- One must be aware that users (p6m) will have changed „their" brands to a certain degree soon afterwards. (In most product categories the „relevant set" consists of approx. 3-4 brands on average. That means, by chance, at least 25 % of the recent users will be lost soon.)

- As a consequence, separate analyses for users vs. non-users must segregate people with an **affinity towards the brand** from non-(interested-) users.

1.2.3.) Households and/or Individuals?

We started our panel operation 25 years ago with tests for *household* products. Consequently, we acquired the co-operation of households, personalized by the **housewife** (=female head of household, reponsible for most of the purchases). Most of our surveys and tests today are addressed to her.

But soon, clients also wanted to test products for men, e. g. shaving foams, deodorants for men, razor blades. And then, special varieties for younger women, girls in their teens; candy bars for children. Consequently, we have included relevant questions to be answered by other members of the household, or had to run special pre-selection postcard surveys to find such target groups.

Last not least cigarettes: For a couple of years we have established a special cigarette-smokers panel. Although most of the women and men in it are being members of panel households, the data file is separate.

Our general experience has been that co-operation by special target groups is practically as good as by housewives.

If the Basis Data Questionnaire shall collect data from the 'housewife' as well as from other members of the family, we strongly recommend to begin the different parts of the questionnaire with clear hints about the person who is to answer this section:

- „the following questions should be completed by .."
- **or:** „...please tick the products you bought within the past 6 months **for yourself"**
- and: „...the following questions are meant for **men**. Please hand out the questionnaire to your **partner** and ask him to answer these questions."

Other products like toilet-rolls or household cleaners are used by the household in general. In such cases one has to decide who has to be defined as „user" and then has to inform the panel-household, who has to answer the corresponding questions (most often the „housewife", who is responsible for most of the purchases).

1.3.) Data Privacy

The German law on data privacy requires that all data should be kept strictly confidential.

Potential panel members should be made aware of the fact that the data collected in the background data surveys are to be stored in the Panel Data File. They will understand the reason for that: so that households will get only products which they use, e.g. non-smokers could be excluded from cigarette tests, tea drinkers from coffee tests, women who always go to the hairdresser's from hair spray tests, and so on.

One major consequence of this law is that the data file has to be separated strictly from the address file. Only by their identity numbers and with the aid of a specific program one can find names of households selected from the data file.

A second major requirement is never to put names and addresses on any questionnaire. We put a sticker with nothing but the household number onto the questionnaire. And a sticker with name, address and number is put on the „window" of the envelope, so that numbers can be compared.

2.) Main Steps in Carrying Out a Product Test

We restrict the description of procedures to product testing; experienced researchers will have no difficulties transferring the procedures to other kinds of tests or surveys.

2.1.) Test Design

In practice, two different testing procedures are being used - plus a melange of them:

1. monadic tests;
2. paired comparisons;
3. sequential tests.

These are discussed in the following paragraphs.

1. Monadic Tests

As the name implies, only one product version is to be tried by one participating household or person, and afterwards characterized and judged with the help of a questionnaire.

The theoretical basis for this kind of test is that under normal circumstances consumers/users would consume/use one product only, not two or more at the same time. This test, therefore, is relatively close to reality.

Consumers/users, of course, judge the test product on the basis of their experiences with existing products, including the brand they used before they had received the test product. And sometimes they are specifically asked to compare their normally used product with the test product. But basically, one is interested to learn the reactions to one product only.

More often than not clients want to test *2, 3, ... n* different varieties of a product. In such a case one has to set up correspondingly *2, 3, ... n* samples, each one to test one version. (These samples may have to be „matched", i.e. they must have the same characteristics. Or, if products are aimed at different target groups, their compositions may vary according to given quotas.

The advantages of monadic tests are:

- Realism;
- when several or many versions of a product are tested (simultaneously or in the course of time), all results can be compared.

The disadvantages:

- Very small differences may sometimes not be reflected, especially if small samples are used;
- larger samples might be necessary than in paired comparisons, and that is more expensive.

2. Paired Comparisons

Test participants get two products which they are to try and then compare. Usually, they get first one product for a given period, then the other one; after the trial period a questionnaire is sent to them which asks for comparisons (sometimes: at first single judgements, then comparisons).

Of course, the sample has to be split into two matched sub-samples: One gets the test products in the order A-B, the other one in the order B-A, the purpose being to neutralize positional effects as good as possible.

The advantages of this design are:

- Very small differences are likely to be reflected;
- one needs only one sample for two products (though 3 mailings instead of 2), which reduces costs as compared with monadic tests.

The disadvantages:

- The trial situation is atypical; reactions are more likely to be quite conscious and rational, thereby becoming unrealistic;
- the „frame of reference“ for each test product is set by the other one; with different pairs in other samples one is likely to encounter differences which are difficult to interpret.

The latter problem has led to the adoption of the „Round Robin“-design, where (theoretically) each test product is tested against every other one. This may solve the problem but can lead to technical difficulties, when many product versions are to be tested: too many and rather small (hard to match) sub-samples are required. And this design does not produce „absolute“ results which could be compared with prior or later tests which are based on different product versions[2].

3. Sequential Monadic + Paired Comparisons

This is a combination of both designs; participants obtain the first test product, then a questionnaire on that. This is the monadic stage.

[2] This was addressed in our paper presented at the ESOMAR Seminar 1996 in Amsterdam, s.a. later-on in this reader

After that they get the second product, which is followed in due time by a questionnaire which asks for comparisons of both products. (Again, one has to use two sub-samples which get the products in the **A-B**, and **B-A** orders.)

This combination is selected with the hope to obtain the advantages of both designs:

- The realism of the monadic tests
- and the ability of the paired comparison test to reflect very small product differences.

For these potential advantages one has to pay:

- One extra mailing costs more money;
- one has two questionnaires instead of one; and a more voluminous report.

Another remark is worth being considered: If results of the monadic and the comparative stages agree, that is fine, - but a waste of (some) money. If results don't agree (and that is not a rare occasion) the discrepancies may not be reconcilable, - and then one doesn't know which one to believe.

2.2.) Sample Selection

The client has to define the target group(s) to be used in the test; or client and research institute discuss alternatives and agree on the definition. Then the institute has to draw a sample or several samples of households or individuals.

Most often matched samples of *„users of ..(category), aged 20-59, with skin type..., representing the proportion of main brands as follows... spread over Germany"* is requested.

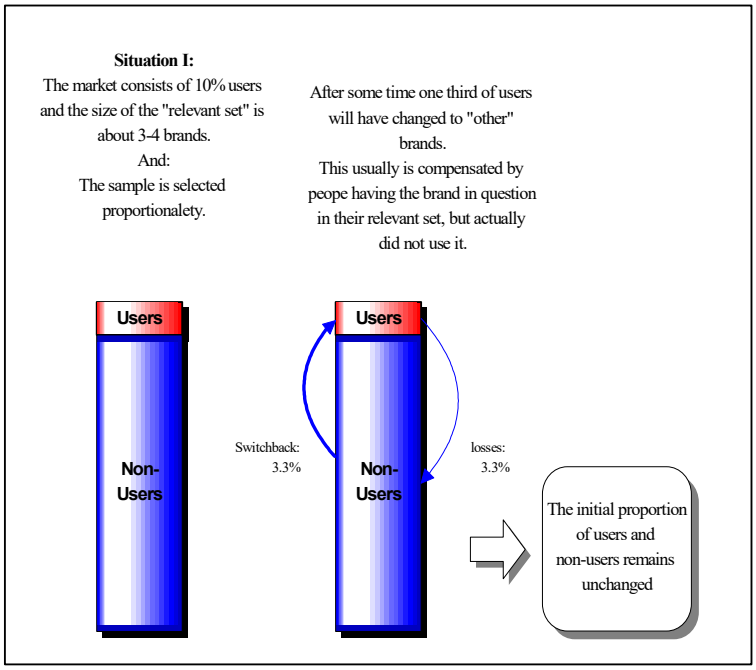
The main point in this request is the question of how to define „brand usership". As emphazised above, the Panel Data File stores an information on „usage (better: purchasing) of brands within past 6 months". This means, **users** can only be defined as people who have the particular brand in their „relevant set". Experience shows shat such people with a certain affinity towards the brand react differently from test participants who wouldn't at all consider buying this brand.

The use of this usership information in the practical process of drawing a sample must consider that the basic question leads to „multiple answers", and consequently, that the users of different brands overlap. But distinct sub-samples of users and non-users request separate sub-groups. Thus we define **users** as those who used the brand within the given period, - if even among others. And **non-users** are defined as those panel members who did **not use** the particular brand at all.

One has to consider the consequences of this definition: If there is a rather long time between the collection and storage of the basic data and the test, the non-users most probably nevertheless will remain non-users (people who did not use a brand within the **past 6 months** presumably will not use it within the next few months).

The problem will be with the „users". As discussed above, one must be aware that those test participants will have changed their brands in an unknown way, particularly, if the brand in question was (6 months ago) *one* brand - and perhaps not the most important one - *among others*. Experience also shows that for larger brands, the probability that „selected users" are also „actual users" in the subsequent test is high.

The opposite emerges if the brand(s) in question are small, and if additionally, a disproportional sample design is requested (50% users, 50% non-users). Here losses must be calculated, as visible from the following picture:

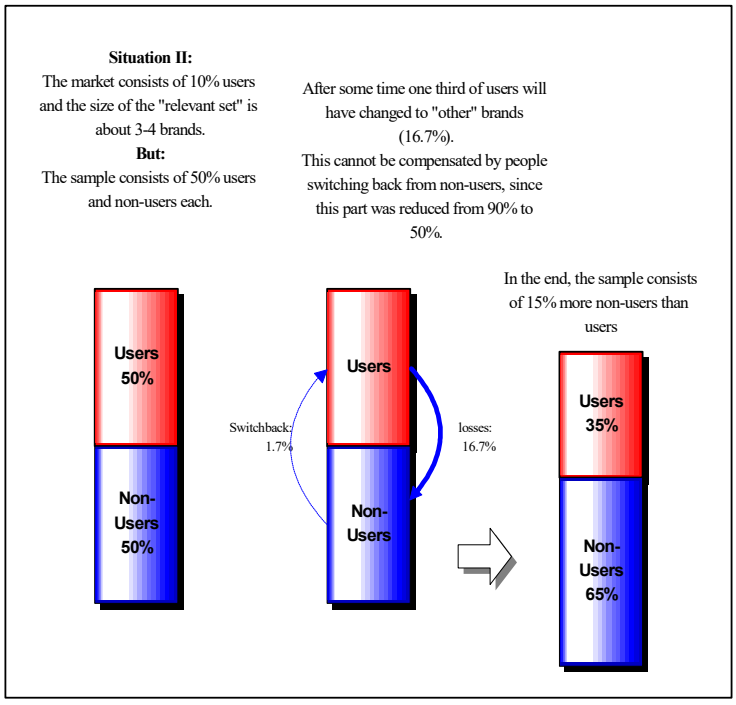


In most cases the data in our Panel Data File fulfil the quota requirements for the individual test; thus we can draw the sample(s) directly from that file.

In some cases we have in the file the data on product category use (e. g. users of diswashing detergents) but not on the brands used; but the latter information is contained in the questionnaires. If required, a sample or all questionnaires are looked through for users of those brands which the client regards as his target group(s).

If the data available are insufficient to meet the client's quotas, a special quick survey has to be carried out, often by post card or by phone. Such a pre-selection survey may be based on a cross section of the panel or on a sample selected according to relevant quotas.

We use an EDP program which can draw samples on the basis of a great number of characteristics. In most cases, two or more products are to be tested; and that requires two or more samples which have identical structures, - so-called „matched samples". Our program can draw such matched samples at random within the quotas set.



Depending on the return rates in previous tests of that particular kind and/or a particular product, the samples drawn have to be somewhat larger than the net samples desired.

The sizes of samples, of course, depend on a number of considerations: Desired accuracy, necessary/desirable splits into sub-groups, cost, and availability of test products (not rarely, the costs of producing test samples is more expensive than the whole test).

Product tests by mail in panels are much cheaper than those carried out by interviewers. And the economy of scale of mail tests is enormous. Consequently, clients find it easier to ask and pay for larger samples and/ or more matched samples so that they may test more versions of the product.

Our samples range between (net) 100 and 400 cases, with the majority between 150 and 200 per test group.

2.3.) Trial

Products are being sent to selected households/individuals accompanied by a letter, which may include an adequate description of the product, if necessary, plus instructions for the use/handling/consumption of it.

Since the public has become very conscious of health and ecological hazards, it is increasingly necessary and advisable to dispatch a full declaration of the product, its contents, and its handling. Whatever is required by laws and regulations to be written on the package of that particular kind of product has to appear on the test product, too, - be it presented „identified" („as marketed") or „blind"! (On principle, product liability rests with the producer. But if it turns out that a producer is in another country and cannot be reached, then the research institute could be held liable.)

In certain cases some kind of concept leaflet is sent together with the test product, so as to explain/promote the special characteristics and specific benefits of it. We recommend to print such 'concepts' on separate leaflets and not on the parallel sent covering letter of the institute. The reason: By doing this we avoid to promote the products' benefits under *our* letter head

The trial period should be long enough so that participants get thoroughly acquainted with the product. Consequently, it is desirable for them to e. g. have completely eaten a box of candies, have completely used up a tube of toothpaste. But the time pressure on the client's side quite often leads to a reduction of the trial period; it should be resisted as much as possible.

Ordinarily, the questionnaire is not mailed with the product. If test participants read the questionnaire before/while using the product it may make them conscious of aspects which may never have occurred to them; and thereby their reactions may be influenced.

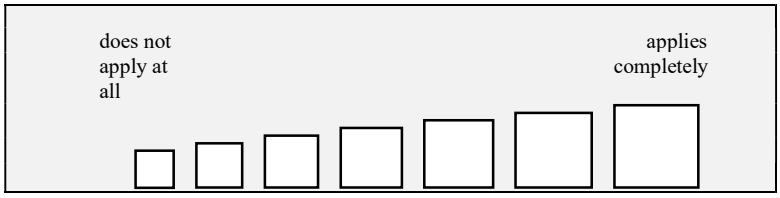
2.4.) Questionnaire

The usual questionnaire for assessing the strengths and possible weaknesses of the tested products consists of the following parts:

- Open-ended questions on *likes* and *dislikes*; usually we start with the latter, asking: *„Any new product should be at least as good as the products you are using usually. Thus: is there anything about the tested product which you did not like so much, anything that could be improved ?"*

  The reason is: R&D is interested mostly in hints indicating sources of future criticism or possible complaints. *Likes* are more for the files.

- Descriptions of product characteristics like colour, consistency, viscosity, lather, scent and taste intensity or duration in terms of *just right,... too much...* or *too little*.
- An assessment of main product characteristics and benefits with the aid of a detailed list of statements, mostly rated on the basis of a 7-pts-scale as the following one:



- A comparison with the usually used product (General Preference), as well as an overall Quality Rating and a scaled rating of the probable Buying Intention.
- Sometimes: price acceptance (following the *Gabor & Grange* approach);
- Target group information & statistics.

2.5.) Coding

Answers on closed questions and scaled ratings generally are transferred from questionnaires to the data file without intermediate coding as soon as the interviews return. The general experience is that within one week after dispatching the questionnaires, 70% returns are available. Then a reminder post card is mailed - leading to a 80% return rate one week later.

Due to the fact that answers to the closed questions and scaled ratings are keypunched in parallel to the returns, we are able to provide our clients ***topline-***results 2-3 days after end of the 2-weeks field-work period.

But answers to open-ended questions have to be categorized first, and that requires experienced coders and additional time. This is nothing new or special. However, there is one aspect which requires consideration:

- Most of our clients are testing products on an almost continuous basis. And they want to compare results over a fairly long time period and for all countries. Consequently, they (and sometimes we) have developed a certain system of coding categories for open questions such as those on „likes“ and „dislikes“. Every research institute participating in an international test has to stick to them as strictly as possible.
- And there we are likely to face even greater difficulties in international studies than we do nationally: Product qualities change, so do the reactions of consumers, - and the latter may be quite different from one country to the next. It would require close co-operation to come to optimal solutions in such cases.

In the meantime we are offering another kind of analysis of the typical open-ended questions. We are calling them *Qualitative Insights*©. Mail surveys benefit from the fact that the respondents' answers are their own words, without being filtered, shortened or pre-coded in an uncontrolled way by interviewers. Our idea was to provide clients just these original answers, without shortening them to more or less uninteresting „overcodes“. With that *Qualitative In-*

sights© provide a deeper insight in what test participants really criticize with regard to the products and by the way, help R&D to find hints for further improvements.

2.6.) Tabulation & Reporting

Most often one institute in an international joint venture is responsible for the final tabulation, analyses and reporting. Thus the code frame and data file description must be given by this institute, and the other ones should stick as closely as possible to this requirements. If any changes are done in either country due to local needs and procedures, this must be well-known to, and approved by, the leading institute.

In our case, we are working with the QUANTUM program package on the basis of its underlying very specific column binary format, - but can also use ASCII files (or others). Meanwhile any transmission of data from one country to another can easily be done via e-mail.

Even if one institute is responsible for the central tabulation, we highly recommend that every institute participating in an international test makes a print-out of its own results, and to check them for accuracy before transmitting the data.

Most clients only want tables with questions (or subjects) and answer categories, percentages and means (for scales), plus certain indications of statistical significance of differences. We are using mostly 3 levels of significance in parallel in our tables, indicated by the following signs:

\+ = significant on at least 1-sigma level
\# = significant on at least 80% level
* = significant on at least 90% level.

The first level indicates a slight tendency that products might presumably not work „equally (well)“. This is particularly important for „cost saving“ problems, where a new product should not be worse than the current one; otherwise it cannot be considered for introduction.

As well-known to any statistician, statistical tests cannot indicate „parity“, but only show „differences“. This 1-sigma level provides a first hint that a difference can be assumed to go in the wrong direction (If a cheaper product is really statistically worse, then it will not be considered in any way. The range between this „significant loss“ and anything that is „close to parity“ is the important one).

The 80% level is then indicating a stronger tendency; and the 90% level finally shows a statistically proven difference. Due to the fact that we calculate these ratings on the base of a two-tailed test, the results indicate in parallel the 90% and 95% levels of one-tailed tests.

Since knowledge of statistically proven differences is only the one side of the coin, we usually conduct factor analyses plus multiple regression analyses in order to ascertain the patterns of relevant product dimension, and to demonstrate their relative importance for evoking an overall impression of quality or even buying propensity.

Some clients want a personal presentation and a written elaboration of results. Samples of typical presentations and reports are available on demand.