

Felix Schaefer:

In-home use tests – A valid method for measuring consumer acceptance in the Health & Beauty Care market

Product tests are one of the most widespread and common tests deployed in the area of consumer goods research. The author has specialised in such tests for more than 25 years and has performed several thousand of them over the years for a range of major, medium-sized and smaller manufacturers of consumer goods, commissioned by their Market Research and Marketing Departments and increasingly also by their R&D and Quality Control departments. The following discussion of potential applications, requirements, methods and case studies is based on a presentation given by us on this issue at a number of universities. But what was interesting and often new to young academics may also be of interest to practical users in the industry – be they from the field of R&D, Marketing or Quality Control.

Possible applications for product tests

Product tests have a whole range of typical applications. One of these is in the **development of new products**, where product tests are used to examine the marketability of a new type of product in the earliest phases of product development; because anything that future consumers do not like or that is not at least as good as what is already on the market is unlikely to have a long-term chance of surviving.

Example 1: As everyone knows, shampoos and hair conditioners have been on the market for some time. Then, one fine day, along came a manufacturer who had the idea of launching a “two-in-one” product, which was intended to unite the features of both products in a single one. The question faced by such a manufacturer and one which can (and could then) be answered by means of a product test is: How well does the two-in-one product perform compared with the established functional competitors? Is it a bit of both, or more one than the other? What benefits are envisaged and what disadvantages?

Such a fundamental question is not the only one that could be answered at an early stage using product tests. Often a whole range of alternative developments are available for the new type of product – such as different consistencies or fragrances. Which one(s) would be the best, which other ones would be acceptable? Product tests are useful here when it comes to making the selection. And when we keep reading about “flop rates” of 70% and suchlike in the marketing literature, this never really applies to those companies that have thoroughly checked out their products before launching them: because those 70% don’t pass the tests in the first place!

The early phase of product development is not only concerned with the best conceivable product, however, but also involves establishing the **competitiveness** of such a new product compared with already established competitors: this can be the market leader or specific, currently successful brands or else representatives of specific market segments. Needless to say, this depends on the direction in which the product launch is aiming. It is not as though every new product needs to be better than Ellen Betrix’ Facial Cream or NIVEA Body Milk and immediately attract just as many consumers – what is necessary, though, is for everything to be “right” in the intended “niche”.

Product tests do not always have to involve new products, though. Examining the **market position of established brands** is another a classic task of product tests, especially with the aim of detecting weaknesses compared with successful competitors. Accordingly, it makes sense to regularly compare the efficiency and acceptance of one’s own brand with that of the market leader, traditional competitors or – in urgent cases – with current predatory competitors.

If a test result has indicated that the quality of the competitor products is better than one's own product, and one's response has been to develop various modifications, then these modifications should also be tested in order to determine whether these newly developed products are capable of successfully replacing the existing product, i.e. the following specific comparisons should be made

- possibly a range of newly-developed products;
- the best of these modified versions with the existing version ("current");
- and the modified and hopefully improved product with the (formerly more successful) competitor product(s).

Quality control is another broad field of application for product tests. We primarily perform such quality control tests on natural products such as coffee, cigarettes, fruit or vegetables – because such products repeatedly give rise to the question: "Is the flavour still right when certain countries of origin have been omitted, owing to crop failures, for example?"

Even in the case of products such as the latter two, which are not subject to any further treatment, product tests can help to determine fluctuations or losses in terms of quality. Ideally, this will take place within the framework of a quality assurance programme such as that described by the pioneer of Japanese quality assurance, the famous professor of statistics W. Edwards Deming (1900-1993) in his book "Out of the Crisis", for example.

He recommends that **all** manufacturers of products and services should subject production processes and even the level of customer satisfaction to systematic, on-going statistical checks. Only then, he claims, can conceivable weaknesses be detected and counteracted in time (and the success of his teachings in Japan is legendary). As far back as the early 1950s in Japan, he committed himself to promoting a philosophy of quality assurance as an "on-going process" – believing that more money could be earned with good quality than by constantly trying to cut costs. His book is definitely worth reading! Americans "rediscovered" him in the 1980s and I believe it was only after his death that the first American company won the "Deming Prize", which was established in his honour in Japan.

What can be tested by means of product tests?

The wide range of applications associated with product tests becomes apparent when you consider all the things that can be examined using such tests: e.g. the acceptance of fundamental components of a product can be tested (i.e. what Myers/Shocker refer to as "*characteristics*" in their remarkable essay entitled "The Nature of Product-Related Attributes"), such as

- scent / odour
- flavour
- consistency / foaming properties
- appearance / colour
- convenience / practicality.

Then the efficiency (or "*benefits*") of products is/are examined:

- basic benefits (effectiveness)
- additional benefits
- or special promised effects.

But we can go even further. Thus the idea/concept behind a product and the presentation thereof, the product name (shower gel vs. after-shave), the brand name, product features or product claims as well as the advertising concept can all be checked.

The various assessments carried out in a product test are usually summarised by determining consumers' general impressions regarding quality, the overall appeal of the product as well as the buying intention and what is referred to as "value for money".

This may sound very obvious, but it has not always been part of general practice – and is sometimes considered controversial even today. When we started testing washing-up liquids for Colgate-Palmolive 40 years ago, none of the chemists working there at the time was able to imagine that "normal" female consumers would be able to distinguish different concentrations of active washing agents and asked themselves – and us – what on earth was the point of doing such tests.

Today, it is quite natural for us to test the effects of various formulas/recipes in a wide variety of products in terms of the perception of relevant benefits, and we ask consumers about the following features, or else have them rated:

- makes skin soft and silky
- provides perceptible long-term care
- or: gives hair body and volume.

This means that R&D develops products and experiments with various caring ingredients, for example (whether Q10 or soya oil). The tests then show whether these various agents and possibly even various concentrations trigger the sense of an "effect" in consumers. And just how far this sense goes. Is there just a feeling that the product "supplies the skin with moisture", that it has a "lasting effect", or does it go even "deeper" ("prevents the formation of lines and wrinkles")?

The same product can display greater or less viscosity. Again the question is: What effect does this have? How do consumers rate its skin feeling, mildness, effect? And finally: fragrance often plays a decisive role too. Depending on whether it is more "unobtrusive" or "fresh", the nature of such freshness – a certain type of caring impression can either be reinforced, or ruined. A day cream for dry and sensitive skin simply must smell differently to an after-shave balsam.

At the same time, testing fragrances – possibly even "pure fragrances" as in the case of "fine fragrances" – is a very difficult task. Neurophysiological and sensory research has made us aware of the fact that although "the human system" is very good at remembering and recognising fragrances, it has few words other than "pleasant" or "fresh" to adequately describe such fragrances. Hence 20 years ago perfumers were just as dismissive about fragrance tests as CP chemist were about washing-up liquid tests 20 years earlier. However, looking at the discussions at the relevant ESOMAR seminars over the past few years it emerges that this too has become generally accepted practice (clashes between sensory experts and market researchers are rare indeed).

Test methods

Most of our tests are **monadic** in design, i.e. they follow the principle *test group vs. control group*. This means the "input" consists of alternative products which are varied in a known or systematic manner.

Each target individual receives *one* product only. The rating of the product is a result of the respondent's experience during the test with exactly *this* product. If certain boundary conditions are controlled in this process, the test automatically supplies a rating of exactly this product – *ceteris paribus*, i.e. under the conditions of the status quo.

The current product range and brands used form the “frame of reference” according to which the test products are assessed. This ensures that the current desires and requirements of test participants, previous experiences and the boundary conditions governing the respective “test situation” are incorporated in the test.

The prerequisites for performing proper monadic tests include:

- random samples of the same structure (e.g. in terms of age and gender but also as regards product use and brand use (= “experience”))
- and an adequate test situation which resembles the typical usage of the test product in question, in particular a sufficiently long test period during which the products can be tested in the usual manner, in sufficient quantities and in typical situations.

Paired comparison tests are often performed as an alternative, whereby each target person is given the test product *and* the comparison product (simultaneously or one at a time). However, here the direct comparison of the given products dominates the ratings of the products, so that the comparative evaluation is carried out by the participants themselves, rather than in the course of the analysis, as in the case of monadic tests.

The key characteristics of these two test approaches are as follows:

a) monadic tests:

- realistic
- benchmark in line with the market
- no obligation to reach a decision

b) paired comparison tests:

- stronger sensitivity
- uniform benchmark (frame of reference)
- lower statistical margin of error (but more test groups required).

These two test approaches can also be combined, and such so-called semi-monadic tests are in fact often carried out.

“Blind” or “identified”?

Another methodological issue is the question of “blind or identified?”

In the case of “**blind**” tests, the products are given to test participants in a neutral packaging with a neutral design. The actual product performance is rated. Blind tests are always performed when you want to know “exactly” what is going on and be sure that no risks are being taken, e.g.

- during the early phase of developing new products
- for product modifications

- in the case of “cost-saving” problems, i.e. when the effects of measures to reduce product costs need to be analysed.

“**Identified**” tests involve test participants receiving the products “as marketed” – often along with a “product concept” in the case of new products. Both the product performance and brand image are rated – just as later on in the real market.

Identified tests are used for new launches when the market presence needs to be analysed and especially also in the case of product improvements, where there is a desire to know whether this improvement will be noticed despite the influence of the image, and whether it will “penetrate” beyond this image, so to speak.

Product input

Where possible, product differences should only be varied “one-dimensionally”, i.e. only as regards the formula *or* fragrance *or* consistency. Interactions must be expected when several variables are altered simultaneously. However a suitable “experimental design” does allow such interactions to be identified through an analysis of the variance, and the influence of individual variables on acceptance can be measured accurately¹.

Survey measuring methods

Product tests resort to all the usual types of questions, i.e. open questions, closed questions and rating scales. **Open** questions are always used

- when the range of components with a possible effect is not yet sufficiently well known;
- when the aim is to discover which impressions have pushed their way into the foreground of people’s awareness;
- when the personal wording of participants is involved;
- when detailed explanations are required.

We have increasingly gone over to evaluating such open questions in the form of a more qualitative “lexical analysis”. Instead of indicating percentages for standard answer categories summarised in the usual way in the form of appropriate classes, we prefer to list the original verbatims to illustrate in full detail the variety of the responses and hence the differences between the various products. Needless to say, mail surveys are very helpful in this respect since they involve the respondents’ actual original answers, which have not been filtered beforehand by the interviewer. We refer to these result as “*Qualitative Insights*”.

Nevertheless, when measuring “hard facts” it is better to use structured (**closed**) questions, particularly when the components that might have a possible effect are already known and it is simply a matter of determining their extent (like it very much, ..., not at all /too much, ..., too little) or else when a measuring criterion can be defined unambiguously in the sense of an exhaustive set of alternative answers (preference questions).

In addition to these, rating **scales** are often a key element of a product test questionnaire, particularly when it comes to communicating gradual differences in product effectiveness (features, functions, attitudes). Verbal scales, numerical scales and optical scales are used here. A classic version of a verbal scale is, for example, the 5-point buying intention scale:

¹ The best-known and most recommendable book on the issue of “Experimental Design” was written by Cochran & Cox.

- I would definitely buy it
- I would probably buy it
- I don't know if I would buy it
- I probably wouldn't buy it
- I definitely wouldn't buy it.

Another widely-used form of the verbal scale is used to rate degrees of intensity, being partly descriptive and partly evaluative, for example

- Is much too strong
- Is a bit too strong
- Is just right
- Is a bit too weak
- Is much too weak.

This type of gradation can refer specifically to a product's sweetness or the intensity of its fragrance, to its consistency or its foaming characteristics. Occasionally, we come across 7-point and even 9-point scales in this context, but in most cases these only lead to monstrous expression such as "extremely much too much ..." etc. and should be avoided at all costs. If a 7-point or 9-point scale is needed, a different type of scale should be chosen (area scale, numerical scale) and verbal anchor points (for each individual level) should be dispensed with.

Item batteries

Our evaluation criteria are always centred around an item battery of variable length describing the essential features of the product, with the help of which respondents are asked to rate the products in question.

Such item batteries should be carefully selected, taking consideration of the various different product features and benefits. Thus the following questions have to be asked each time:

- What features does a shampoo have?
- What features does a body lotion have, or a facial cream
- What are the differences between a day and a night cream?
- What additional benefits do anti-wrinkle creams offer?
- What influence does fragrance have?
- How are fragrances rated – or even perfumes?

Alfred Politz taught us to think a step further at times:

"We noticed that something was wrong with the cigarettes at this time ... and we conducted interviews ... asking them: Why do you like cigarette ... better than ...?"

The answer: "It has more tobacco taste"...

... is wrong; we have no tobacco taste in cigarettes, because the additives, heat treatments, and everything else that goes into the cigarettes distorts the tobacco taste.

However, we built new cigarettes step by step ... and the cigarette that got the most frequent response of "most tobacco taste" was the cigarette we launched ...".

Interview in JAR, June 1977: "The Founding Fathers"

Where possible, we do not use "individual words" for such statement batteries, preferring "statements" instead; nor do we use "bipolar scales" (semantic differential). We have already discussed this issue in some depth at the 2nd "planung+analyse" symposium in connection with the criteria used in advertising media tests.

Multivariate methods

Factor analyses can be used to compress the long and multifaceted list of assessment criteria and to establish "general dimensions". And with the help of multiple regression analyses based on these factors, it is possible to determine just how strongly the individual factors influence general perceptions of quality or the "buying intention". This is important additional information, because while it is of course "nice to know" whether two products are significantly different, the real question is whether this difference is actually *important*.

Example: Facial Day Creams:

Factor 1: Caring Properties & Mildness

- provides skin lastingly with the necessary moisture
- provides skin lastingly with protection from drying out
- has a noticeably long-lasting caring effect
- gives your skin new elasticity
- makes skin smooth and supple
- noticeably improves the appearance of skin
- regulates the skin's natural balance of oils
- is particularly mild and tolerated well by skin
- leaves a pleasant feeling on skin

Factor 2: Consistency & Application

- has a pleasant consistency
- can be spread well on skin
- is absorbed quickly by skin
- is pleasant to use

Factor 3: Characterisation of Scent

- has an unobtrusive scent
- has a fresh scent
- has a pleasant scent.

As we know from our database analyses, the first factor is the most important one for establishing a sense of quality and triggering "willingness to buy the product". This applies relatively equally both in Germany and in France, as can be seen from the following regression coefficients:

| | Germany | France |
|--|---------|--------|
| Factor 1: Caring Properties / Mildness | 0.67 | 0.62 |
| Factor 2: Consistency / Convenience | 0.41 | 0.52 |

Factor 3: Scent

0.27

0.22

This means the “willingness to buy” increases 50% more quickly when the ratings for the first factor are increased, than it does when consistency ratings are improved, and about twice as fast as when fragrance ratings are good or better. Which means in turn that a better overall impression can be generated more quickly and more effectively by “tinkering” with the consistency instead of changing the fragrance.

“In-home use” test

The overall performance of a product can only be realistically tested in the context on an “in-home use” test; here the products are tested in a familiar environment, under typical conditions and at typical frequencies and levels of intensity. This implies a test period which comprises all of the conditions governing typical product usage as well as product quantities that permit typical frequencies and intensities of use.

Our research agency runs a nation-wide product test panel comprising almost 50,000 households which has proved particularly successful as a basis for implementing in-home use tests in particular. The most important structural and usership data is stored and available for selecting target groups. Computer-aided selection guarantees the structural equality of random samples (“*matched samples*”). The particularly high motivation of panel households to cooperate in such tests reduces the already low costs of mail surveys thanks to high return rates.

The method of mail surveys has proved its worth in more than 3,000 tests. Typical prejudices towards mail surveys have long been refuted by our own practical experience and discussed at length in the German market research journal “planung+analyse”, which has published an special series about “mail surveys”.

Findings

Product tests which are conducted by mail in a panel respond extremely sensitively to all kinds of influences which might affect the *input*:

- Differences between products
- Packaging shapes and materials
- Seasonal conditions
- Changes in the target group

This is both a risk and an opportunity. When products are varied in a systematic and a known way and the extraneous influences are controlled, these tests supply the desired answers to the questions posed – i.e. they are an positive aid to decision-making. We presented some typical examples of this at the CSC Conference, e.g. the effects of various qualities of foam on the perception of mildness and skin compatibility:

| | A | B | C |
|------------------------|------------|-----------|------------|
| | % | % | % |
| too much foam | 3 | 3 | 1 |
| just right | 52 | 71 | 83 |
| too little foam | 45 | 36 | 16 |
| foams well | 4.3 | 5.2 | 5.6 |
| has creamy foam | 5.0 | 5.6 | 6.0 |
| has fine foam | 4.9 | 5.3 | 5.8 |

**is particularly mild
and gentle on skin**

5.0 5.8 6.0

Unfortunately however, there are repeatedly cases in which the results are not as expected. Usually, it transpires that the test was in fact “right”. Let me give you some examples:

- In most tests, the differences between products are “desired”, i.e. brought about on purpose. There are also however “undesired” test issues caused by differences in the conditions during production and subsequent storage, for example.

Filling processes performed in technical centres / laboratories can vary and in turn differ from the normal “fresh” production process, as well as from repurchases from the retail sector. Without being aware of all the things that can happen, you would be surprised at how different the results were (and usually start by blaming the test).

- Another example concerns “packaging differences”. Packaging tests involve determining whether and to what extent the packaging is appealing and how it influences perceptions as regards quality – in general and in terms of key benefits.

Accordingly, lotions or shampoos are tested in bottles with and without spray inserts, with the aim of finding out whether this economy measure (each gramme costs money in Germany because of the “Green Dot” – an organisation which collects and recycles plastic packaging) is accepted without impairing quality perceptions (whereby handling also plays a role, of course). Or glass bottles/pots are compared with the corresponding packaging made of plastic.

By coincidence, we once tested exactly the same (blind) product in containers made of various different *plastic materials* and analysed the results according to the type of plastic used. It transpired that these plastic bottles made of different materials and used by coincidence had a direct influence on the ratings of “feeling on skin” and “mildness”, over and above the “tactile” sensation they themselves produced. So be careful when choosing the packaging for your blind tests!

- New market conditions can also alter product test results. For example, the appearance of two-in-one products in the *shower gel* range led to a new segmentation of this market. Before that, there were medicinal shower gels, mild/pH-neutral shower gels, caring shower gels, refreshing and perfumed shower gels as well as all kinds of body lotions. The questions arising after introduction of the two-in-one products included:
 - Where does a two-in-one shower gel and body lotion position itself?
 - What other and whose demands are now met by a *caring* shower gel?
 - What does the target group for the latter look like and what are earlier test results worth now?

The “frame of reference” sometimes changes faster than you would think: people shower and bathe more in winter than in summer; wind and rain in autumn make different demands on hairsprays than during the summer months; when you’re nice and tanned in the summer, you choose different make-up colours to those at other times of the year; people use different fine fragrances on “festive occasions” etc. All of these circumstances can also affect the results of product tests and should be known as influences – or at least one should be aware of them.

Prognostic strength

Product tests primarily provide “statistical comparisons”, i.e. initially they simply indicate whether one product is significantly “better” or “worse” than another. If comparative data or database findings are available, it is possible to deduce whether this is synonymous with “market maturity”. For this reason, we have for many years been maintaining databases in which all results are stored in the form of control charts, enabling us to provide reliable “benchmarks”.

Nevertheless, “ultimate information” can only be provided by special market tests designed to give a prognosis, such as:

- Concept & Use Tests (“*CUTE*®”) or
- Prognosis Tests (“Simulated Test Market”, “ConTesi®”).

Further reading

- Bauer, E.: “Produkttests in der Marketingforschung”, Vandenhoeck & Ruprecht, 1981
Cochran, William G., Cox, Gertrude M.: “Experimental Design”, Wiley & Sons, N.Y., 1957
Deming, W. Edwards: “Out of the Crisis”, MIT Press, 2000
Gruenwald, G.: “New Product Development”, NTC Business Books, 1988
Hardy, Hugh S., (ed.): “The Politz Papers: Science & Truth in Marketing Research”, American Marketing Ass., 1990
Holm, K.-F. (pub.): “Produktforschung” – 3rd pl+a Symposium, M+K Hansa Verlag, 1987
Politz, Alfred: Interview “The Founding Fathers”, Journal of Advertising Research, June 1977
Schaefer, F.: “Kann man Produkttests so erweitern, dass sie zur Prognose von Marktchancen geeignet sind?”, Planung+Analyse, 1983
Schaefer, F.: “Erfahrungen mit einem Prognose-Test”; BVM series, Volume 24, 1995
Urban, G.L., Hauser, J.R.: “Design & Marketing of New Products”, Prentice Hall, 1980
Wind, Y.J.: “Product Policy: Concepts, Methods, and Strategy”, Addison-Wesley, 1982
ESOMAR Seminars on “New Product Development”, 1974, 1979, 1984
ESOMAR Monograph Series, Vol. 1: “New Product Development”, 1988
ESOMAR Seminars on “Research on Flavours & Fragrances”, 1989, 1991, 1993, 1996.