Testing Fragrances & Flavours
Monadically or by Paired Comparison?

by Felix Schaefer, Hamburg

Session I of this seminar describes several experiences with product tests and mappings as development tools. One paper deals in particular with „preference maps from paired comparisons“.

The present paper adds some of our experiences in drawing mappings on the basis of results from paired comparison tests. It shows that the order of presentation does have an influence on the perception and the rating of product characteristics. It also shows that the rank order effect cannot be neutralized properly by switching the order of product presentation in a random half of each sample. As a consequence, mappings based on paired comparison tests are - in our experience - difficult to interpret. Thus, if at all possible, we recommend running monadic tests for such purposes.

The experiences we describe derive from more "solid" products, i.e. roasted coffees, but the same can certainly be found in tests with the more ephemeral characteristics of (fine) fragrances.

The Question of "The Better Method"

The question is quite an old one, the advantages and disadvantages of the two methods have been known and debated to and fro for decades. No wonder that as early as 1966, Blankenship suggested in an article

"Let's bury paired comparisons!"

The reply came swiftly: Haller argued

"Let's not *bury paired comparisons".*

And there, it seems, we still stand today.

The typical arguments regarding paired comparisons are:

pro:
- economy (one sample necessary for testing two products)
- dependability of samples causing smaller variances
- magnifies minor differences.
- simplicity of analysis & interpretation (X percent prefer A, Y percent B)

contra:
- atypical situation regarding product use
- influences due to order effect, since each test establishes a specific "frame of reference"

The arguments regarding monadic tests are:

pro:
- realism
- results can be compared between several test groups and between series of tests/ over time
- additional statistical analyses (factor-/ regression-/ cluster analyses) are doubtless possible

contra:
- larger samples necessary
- at least two samples necessary
- matched samples necessary.

The advocates of "paired comparisons" concede that the order of presentation may have an influence on the results, but argue that this influence can be neutralized by the well-known technique of switching the order in a random half of each sample.

We are able to show that the two orders do not neutralize each other; but that each order produces its own kind of influence. This may be acceptable in isolated tests that are to be analysed separately. However it is hazardous for mappings comprising the results of a series of tests: they do not produce stable, interpretable positions for the mapped products.

Steps in Experience

a) Starting With the Classical Form of Paired Comparison Tests

Our institute has been conducting tests with roasted coffee since the early eighties. The objective of these tests was to improve the quality of roasted coffees step-by-step. In the very beginning no one knew to what degree and how successfully small differences could be measured. Thus, the client asked us to conduct paired comparison tests to reveal even the smallest differences.

We started with the "classical" **paired comparison** approach:

- two different varieties of coffee were dispatched - the "current" product as a benchmark, and a new one;
- the test participants were asked to use them one after another (each for one week);
- one half of the test groups were to use the products in the order A - B, the other half in reverse order B - A;

- at the end of the testing period, test participants were asked the usual questions:
  - „Which one did you like better ?“ (with forced choice)
  - and: „Why ?“

The results were analysed test by test. If the new product was preferred, it was launched, and in subsequent tests it was used as the next benchmark. If the new variety did not win, it was to be improved on the basis of given ad hoc findings - derived from "likes" and "dislikes", respectively "reasons for preference". Then this next new variety was tested using the same procedure.

Apparently this paired comparison test approach was able to fulfil the client's demands for rather a long period.

Sometimes so-called Round-Robin Tests were conducted, in which more than one new product was to be tested, and each alternative was compared with every other one. Anyone who has had some experience with such tests knows that very often such multiple comparisons lead to the surprising result that product "A" is preferred vs. "B", and "B" vs. "C" - but **not** "A" vs. "C". But mostly such results can be "explained" by looking directly at the given results. Consider the following:

Test group 1:	Milder coffee **A**	vs.	stronger coffee B
(vice versa:)	Stronger coffee B	vs.	milder coffee **A**
Test group 2:	Stronger Coffee **A**	vs.	stronger coffee C
(vice versa:)	Stronger coffee C	vs.	stronger coffee **A**.

Apparently, coffee "A" may get 4 different ratings in these alternative comparison situations. Nevertheless, very often the ad hoc given findings of such isolated tests do not cause too many problems for the end user to interpret the results: they are explicable.

b) Using Scaled Ratings as Additional Diagnostics

After some time, 12 statements were derived from the open-ended question on „reasons for preference“, describing the most relevant characteristics and benefits of roasted coffees. Thereafter, respondents were in addition asked to rate each of the tested products with the aid of these relevant statements using a 7-points- scale.

The direct comparison helped - as before - to find the better one of the two products, and the scaled ratings provided the necessary diagnostics on a broader basis than before.

Using scaled ratings has had a positive side-effect: it became possible to run factor and regression analyses to establish patterns within this set of items and to find out their relative importance in affecting overall acceptance and buying intention.

The Factor Analysis produced the following 3 factors: [1]

Factor 1:
- has a strong/rich taste
- has a full aroma/ flavour
- is perceptibly full of caffeine
- is very economical in use

Factor 2:
- is not harmful for the heart
- does not upset stomach
- is easy to digest

Factor 3:
- tastes bitter/sour
- has an unpleasant aftertaste
- tastes (not) mild.

A "Multiple Regression Analysis" was run to rank these factors in the order of their influence on buying intention. Typically this analysis showed that it was more important to improve the product characteristics subsumed under Factor 1 than those under Factor 2, while at the same time keeping the contra-productive characteristics of Factor 3 at a low level:

[1] In the meantime we have added further statements to the above list which provide a much more detailed description of the relevant product characteristics and benefits of roasted coffees; from this list, 6 factors can be derived today.

Buying Intention =	**0.53**	x	results achieved on Factor 1
	+ 0.27	x	results achieved on Factor 2
	- 0.38	x	results achieved on Factor 3.

Consequently, this analysis helped to give further developments a clear direction.

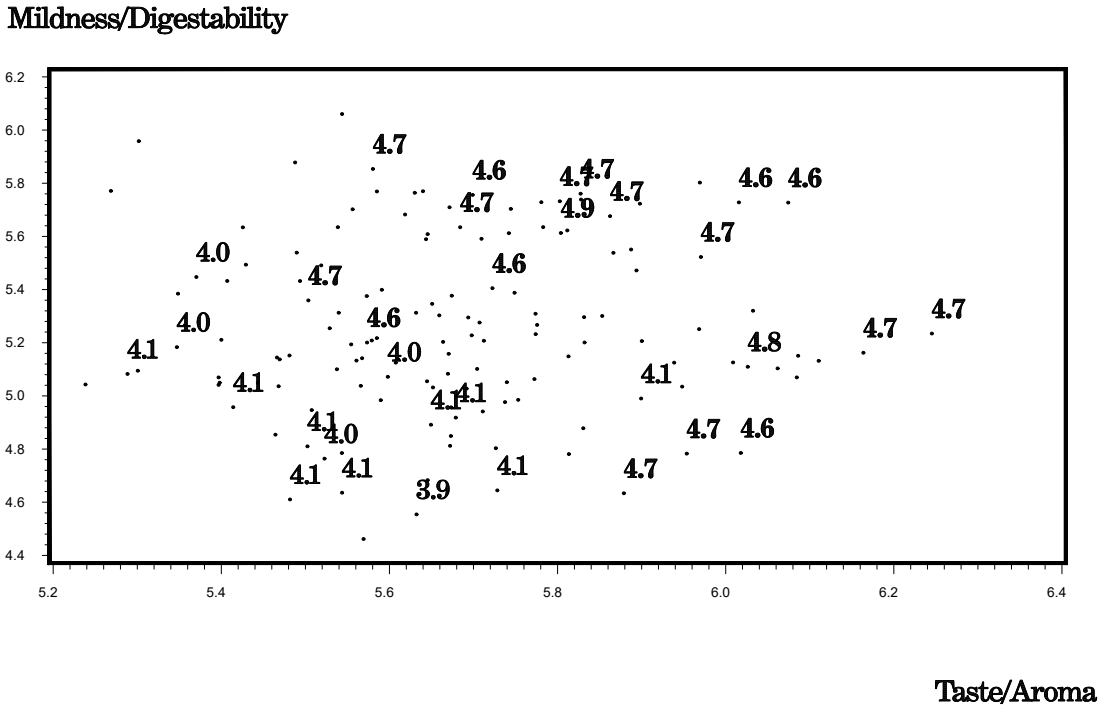
c) Data Base: General Findings & Mappings

Over time, more and more tests were conducted and numerous results from various tests with different types of coffee were available. Most often, the multivariate analyses conducted confirmed the above findings and it was fair to expect that analysing all these data on a broader basis would help to reveal findings of a more general nature.

Thus in 1985 a data base was established, comprising 140 test groups with results for normal coffees, "top" coffees, mild coffees and decaffeinated ones.

And indeed, the overall factor analysis confirmed the above factor pattern, and the subsequent multiple regression analysis revealed the same rank order of the factors' relative importance. The next diagram shows the positions of the tested varieties in relation to the two first and most important factors. In confirmation of the result of the multiple regression analysis it reveals that indeed the coffees that were liked better, gaining the highest buying intention scores, were positioned in the upper right-hand corner of this „map“:

Graphic 1



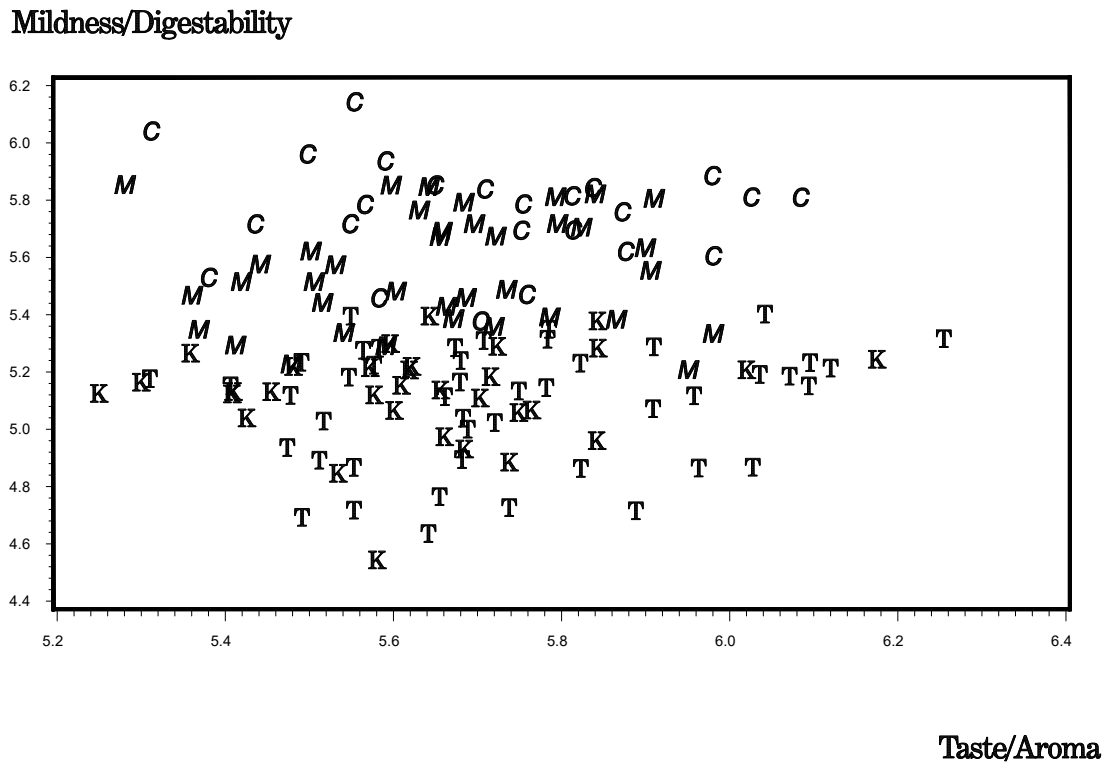
This was good news for the R & D people, since from now on, they had a pretty good guideline for making further improvements: "Try to develop a full-flavoured but mild and digestible coffee and you will be successful !"

The next graph shows a clear differentiation of the different types of coffees at least for Factor 2:

- the mild & decaffeinated coffees (=M/C) are positioned in the upper area of the graph, indicating their higher compatibility/ digestibility,
- whereas the "normal" & "top" (=K/T) coffees are positioned more in the region indicating a lower degree of mildness of such types of coffees.

This also underlines the suggestion that such factor & regression analyses and the subsequent mappings can be of assistance for further product developments.

Graphic 2



Possible Pitfalls of Paired-Comparison-Test Mappings

The problems of mappings based on paired comparison tests become visible, if one looks at the specific positions of those products which were tested more than once.

As described before, the situation may occur where comparing products "A", "B" and "C" does not provide consistent results. But now, this is aggravated by the fact that the numerous statements given describe and allow various "points of view". Comparing a „strong“ product with several "milder" ones may change the individual frame of reference for each comparison, depending on how perceivable differences in the manifold aspects of "strength" are, as e.g.: "full flavour", "strong taste", "harmful for stomach" or "invigorating caffeine content".

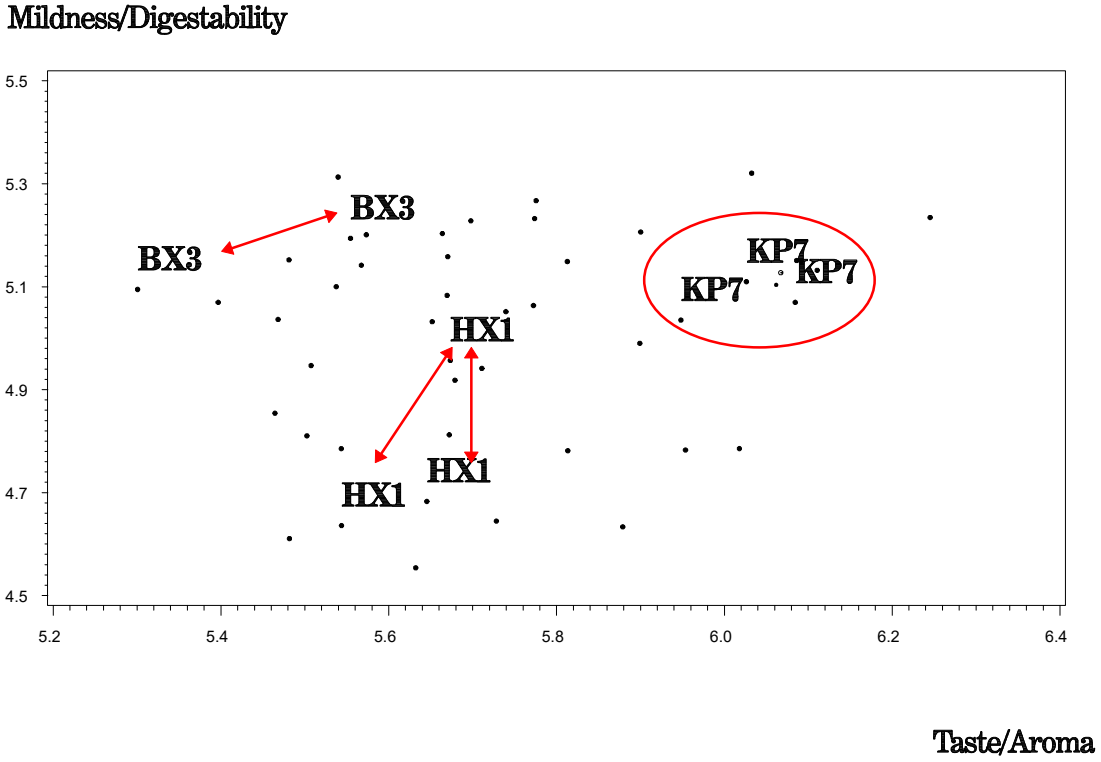
Consequently the position of one and the same product changes in terms of "strength", depending on what was responsible for establishing the underlying frame of reference.

The following diagram (an others were shown at this ESOMAR-seminar) shows several products from the "top" coffee range which obtained clearly different positions when compared with different alternatives. Since the underlying samples were strictly matched, this cannot but being affected by the "frames of references", set by each pair in the mind of the respondents.

Obviously this effect could not be balanced or even cancelled by switching the rank order in one half of each sample. The „frames of references“ clearly dominated the outcome, and analysing such a mapping is virtually worthless:

- which one is the "correct" position of a tested product ?
- what can be done to move the products' position from X to Y ?
- how can we relate objective product parameters to subjective consumer perceptions & judgements?

Graphic 4



None of these interesting questions could be answered adequately. Our recommendation was thus to **„bury paired comparisons“** and to switch to monadic test procedures.

In the meantime, we have established a new data base of *monadic* test results. Its data can be compared over time & seasons, for "blind" and "as marketed" tests, different types of coffees (normal, "top", mild & decaffeinated) and last, but not least - with R&D data.

References:

Axelrod, Joel N.: "Reducing Advertising Failures by Concept Testing", Journal of Marketing, October 1964
Blankenship, Albert B.: "Let' Bury Paired Comparisons", Journal of Advertising Research, March 1966
Greenberg, Allan: "Paired Comparisons vs. Monadic Tests", Journal of Advertising Research, December 1963
Haller, Terry P.: "Let's *Not* Bury Paired Comparisons", Journal of Advertising Research, September 1966
Seaton, Richard: "Why Ratings Are Better Than Comparisons", Journal of Advertising Research, February 1974
Schroiff, Hans-Willi: "All's Well That Smells Well ?", ESMOAR-Seminar on "Fine Fragrances and Fragrances in Consumer Products", London 1991

Shalofsky, Ivor: "Fragrance Choice and Perceptual Mapping", ESMOAR-Seminar on "Research for Flavours and Fragrances", Lyon 1989