

A systematic approach
to planning and performing

Sensory Tests



Choice of the test method and appropriate panel selection criteria

› Maria João Monteiro

Escola Superior de Biotecnologia
Universidade Católica Portuguesa



CATÓLICA
UNIVERSIDADE CATÓLICA PORTUGUESA | PORTO
Escola Superior de Biotecnologia



**European
Sensory
Network**

EUROPEAN Sensory NETWORK
joint partner for sensory & olfactory research

50 years in search of the consumers' true motivations

European Sensory Network co-founder and scientific advisor Egon P. Köster, sensory expert and Professor emeritus of Experimental Psychology, by occasion of his 75th birthday

“People only eat what tastes good to them and only buy what they like. Yet the question of what is accepted by whom is not easy to answer. Consumer sensory research attempts to get to the bottom of this question and to discover which product attributes are decisive and which preferences influence the consumers' purchasing decisions”.



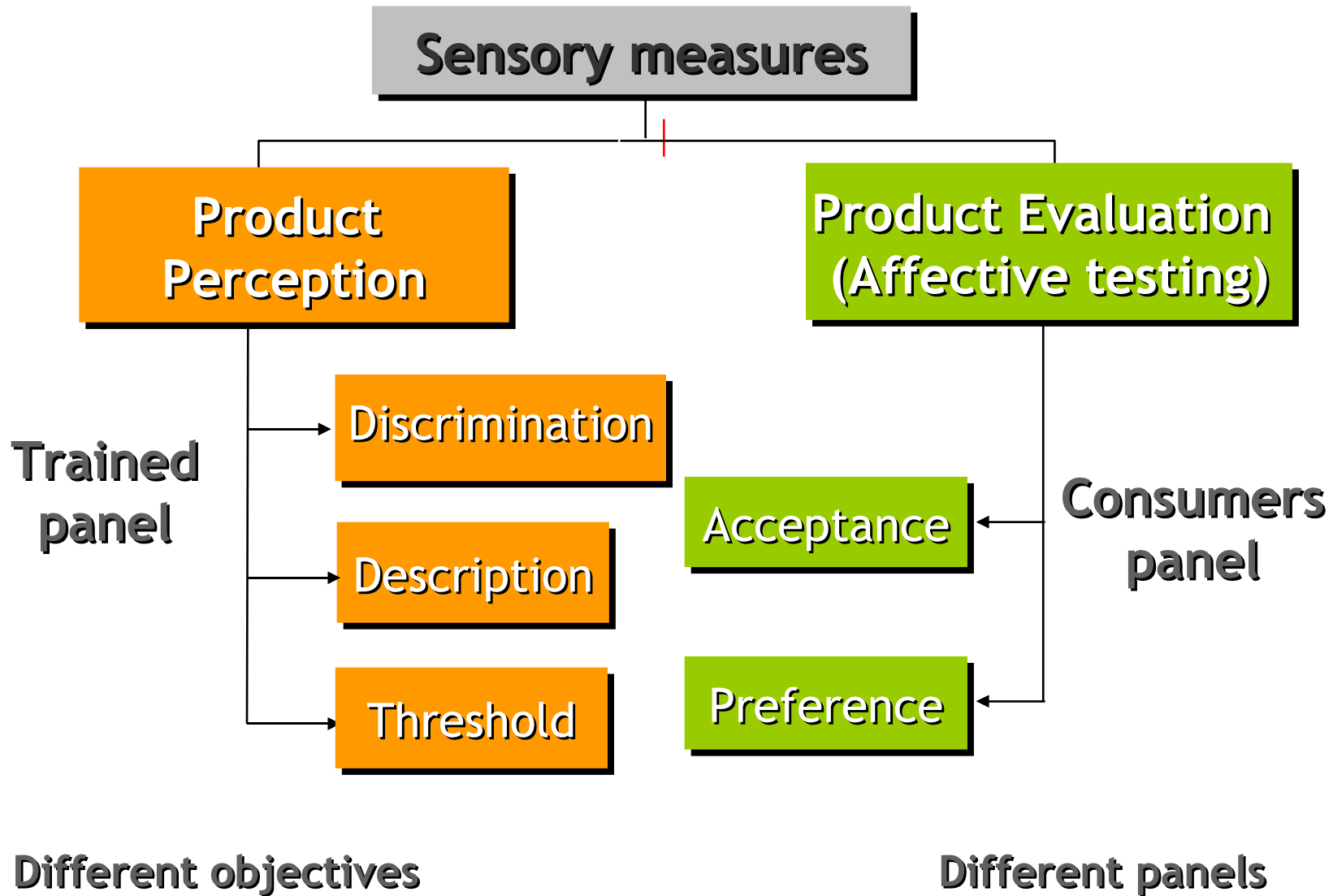
Professor Egon P. Köster, Interview Dec. 2006



***“Measure what is measurable, and
make measurable what is not so.”***

Galileo Galilei (1564–1642)





DISCRIMINATION TESTS

Overall
difference
tests

- Does a sensory difference/similarity exist between samples?

Attribute
difference
tests

- Is there a perceptible difference/similarity between the products concerning the intensity of a selected attribute?

Choice
tests

- Assessors are asked to choose:
Different or same?
More intense, less intense?
Better, worse?

Direct scaling
methods

- Assessors are asked to measure:
Degree of difference?
Degree of liking?
Degree of intensity?



Objectives

Are the products similar or are they different in any way?

Discrimination tests

What are the product sensory attributes, what is their intensity?

Descriptive tests

Are the products liked, which is the preferred?

Affective tests

- **Product**
- **Assessors**
- **Desired level of confidence in the conclusions**



Affective testing

➤ **Consumer panel**

No relevant experience or training
Selection based on representativeness

Product perception

➤ **Trained panelists**

Selection and training: objectivity, precision, accuracy and reproducibility

Trained in test methodology and type of product

➤ **Expert panelists**

Capable of evaluating differences and explaining their causes

The required level of qualification (experience and ability) of assessors has to be carefully considered according on the specific test objective/situation.



CHOICE TESTS

Triangle test



Which is the odd sample?

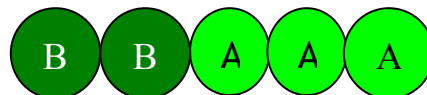
MOST USED

Duo-trio test



Which sample is the same as the reference?

2-out-of-5 test

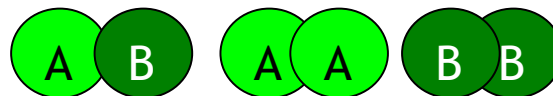


Product presented twice?

SENSORY FATIGUE

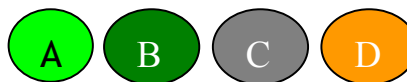
DIFFICULT TO MEMORIZE

Simple difference



Are the samples the same or different?

“A” not “A”



Is the sample “A” or “not A” ?



➤ Features

Applicable only to fairly homogeneous products

Triangular, duo-trio and 2-out-of-five, enable estimation of population's proportion able of detecting a difference

Limited information: determine neither the size nor the direction of difference.

Inadequate for comparison of moderate/large number of products.

➤ Assessors

Level of qualification dependent on the specific test objectives

Panel dimension function of type of test and sensitivity (α, β, p_d)

α	β	p_d	Duo-trio	Triangular	2-out-of-5
0.05	0.30	0.4	30	16	7
0.05	0.05	0.4	67	40	18
0.05	0.05	0.2	268	147	49



Applications

- Selection and training of assessors
- To determine whether a sensory perceivable difference results or not from a change in ingredients or process
- Threshold estimation (triangle test)
- Limited application to Quality Control/Quality Assurance



Quality Control / Quality Assurance

Monitoring product consistency

Monitoring raw material consistency

Monitoring changes in product formulation

Monitoring changes in manufacturing process

Quality Control/Quality Assurance usually based on
“limited variation” rather than “no difference”

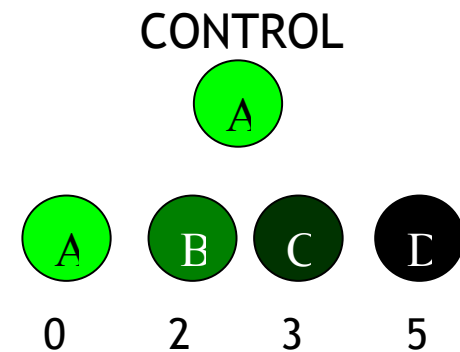
Inadequacy of choice sensory tests ?



SCALING TEST

Difference-from-control

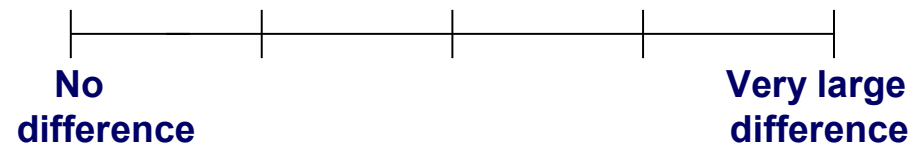
One product is the designated “control”.
All others are evaluated with respect
to the size of difference to the control,
using the provided scale.



Verbal category scale

- No difference
- Very slight difference
- Slight/moderate difference
- Moderate difference
- Moderate/large difference
- Large difference
- Very large difference

Continuous scale



Randomized Complete Block Design or Balanced Incomplete Block Design



> DIFFERENCE-FROM-CONTROL TEST

Assessors:

Level of qualification dependent on the specific test objectives.

Assessors consistency should be checked for more precise results.

Advantages:

Applicable to **heterogeneous** products.

Estimates the **size of difference** to the designated control (but not the direction or the attributes responsible).

Relative size of a variation is important for decision making:

Quality control/assurance

Impact of ingredient, process or package changes

Shelf life studies



Situation:

Two wine tanks. One suffered a bacterial contamination. Want to know if the contamination had sensory perceptible consequences.

Testing: Triangular test,
24 assessors,
14 correct identifications of
the odd sample

Conclusion:

In fact a perceptible difference was observed. At 95% probability, the mean population proportion able to detect a difference was estimated in 38% with a minimum 13%*.

*assuming the population had similar discriminating ability of panel used



Situation:

- One product currently produced;
 - 3 different formulations of breakfast cereals (antioxidants composition) under development.
- Perceptible sensory difference?

Testing conditions:

- Difference-from-control test,
- 23 trained assessors
- continuous scale

0 - no difference

6 - large difference



Sample	Blind control	Formulation A	Formulation B	Formulation C
Mean response	0,5	1,0	1,6*	1,9*

LSD_{95%} : 0,7

* Significant difference 95% confidence level



ATTRIBUTE DIFFERENCE TESTS

Is there a perceptible difference/similarity between the products concerning the intensity of a selected attribute?

Choice tests

- Paired comparison (2-AFC)
- Pairwise ranking test

Measurement tests

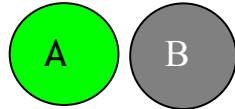
- Grading
- Ranking
- Scaling
 - Magnitude estimation
 - Rating

Hedonic and non-hedonic testing

CRITICAL: the attribute under test must be clearly defined and understood by assessors



Paired comparison

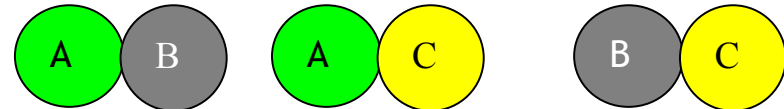


Two products,
two samples presented

- In which sample is attribute X more intense?

Very easy to perform and easy to interpret
Very limited information

Pairwise ranking test



Comparing several samples in all possible pairs

Nº samples	2	3	4	5	6
Nº pairs	1	2	6	10	15

- Which sample do you prefer?



**Ordinal scales
(unequal intervals)**

**Interval scales
(equal intervals)**

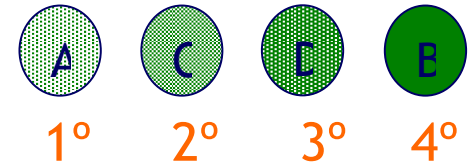
**Ratio scales
(proportional scales)**

- Ordinal scales require non-parametric statistical analysis. Non-parametric statistics are more robust than parametric ones (less affected by anomalies in data) but are usually less powerful than parametric tests (if a difference exists, the parametric test will be more sensitive in demonstrating it).



RANKING TEST

Ranking samples according to the intensity of designated attribute or preference (ordinal scale)



- Easier and less fatiguing than other measurement methods
- Moderate statistical skills required for data analysis

Main applications:

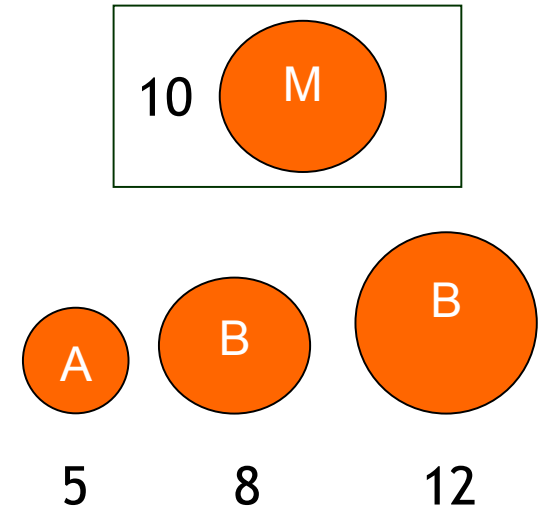
- panel training and assessment
- pre-sorting a large number of samples in product development and consumer testing

Main disadvantages: Intensity of attribute and size of difference between samples is not determined



MAGNITUDE ESTIMATION

1. A value is assigned to the intensity of the attribute of interest of the first sample.
2. Subsequent samples are rated in **proportion** to the first sample.



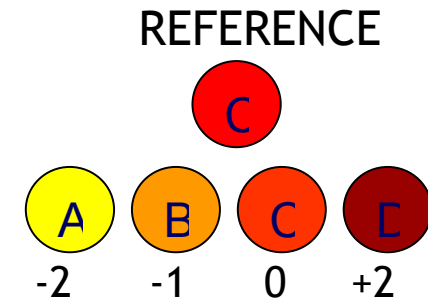
- Requires specific training of assessors
- Not adequate for differentiating small differences
- Not adequate for evaluation of intensities near threshold level
- Does not provide absolute ratings of the intensity of the attribute

Classifications are not affected by scale end-effect



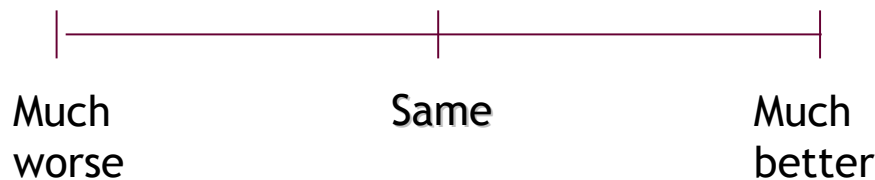
ATTRIBUTE - DIFFERENCE-FROM-CONTROL TEST

One product is the designated “control”.
All others are evaluated with respect
to the size of difference of the assigned
attribute, to the control, using the
provided scale.



Stronger	Better
Same	Same
Weaker	Worse

Overall liking



Attribute X

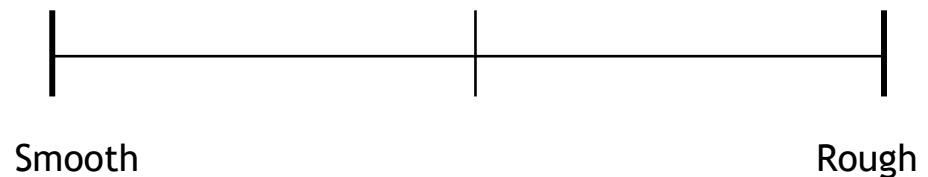
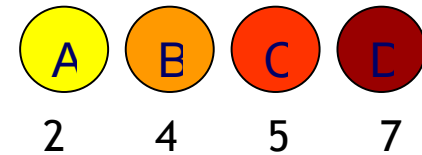
- +3 Much stronger
- +2
- +1
- 0 Equal
- 1
- 2
- 3 Much weaker



15	Strong
12,5	Moderate-strong
10	Moderate
7,5	Slight-moderate
5	Slight
2,5	Very slight
1	Threshold
0	None

6	Very intense
5	
4	Intense
3	
2	Weak
1	
0	Absent

The intensity of the selected attribute is rated on a specified intensity scale



Classification scales:

- Broad enough to include full range of parameter intensities;
- Broad enough to enable discrimination of small differences;
- Panelists tend to avoid the extremes of the scale, distorting the scale. In discrete scales, the smaller the number of categories, the greater the end-effect.
- Intensity can usually be more accurately graded with line scales. However, less trained assessors usually consider line scales more difficult to use.



Assessors:

- Results are critically dependent on assessors qualification and consistency:
 - Training should include familiarization with the range of products and the use of classification standards.
 - Consistency periodic checking

Applications:

- SCALING tests largely used in quality control/assurance, product development, ingredient, process or package changes evaluation and shelf life studies.



PLEASURE EVALUATION

9 Point Category Hedonic Scale

9	Like extremely
8	Like very much
7	Like moderately
6	Like slightly
5	Neither like nor dislike
4	Dislike slightly
3	Dislike moderately
2	Dislike very much
1	Dislike extremely

PERYAM and GIRARDOT 1952;
PERYAM and PILGRIM 1957

**The most widely used scale
to determine consumer acceptance**

Labelled Affective Magnitude Scale LAM



SCHUTZ and CARDELLO, 2001



“Unfortunately, 9-point Hedonic scale suffers from problems related to unequal scale intervals and the under use of end categories.

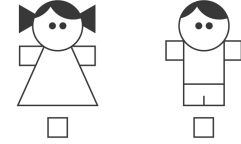
LAM scale was shown to have equal reliability and sensitivity to the hedonic scale, provided somewhat greater discrimination among highly liked foods, and resulted in data that were similar to magnitude estimation in terms of the obtained ratios among rated stimuli.

The LAM scale was also judged by consumers to be as easy to use as the 9-pt hedonic scale and significantly less difficult than magnitude estimation.”


SCHUTZ and CARDELLO
Journal of Sensory Studies, 2004



> HEDONIC RATING



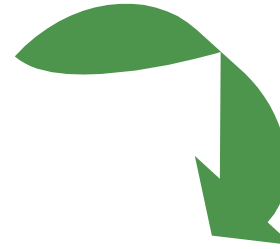
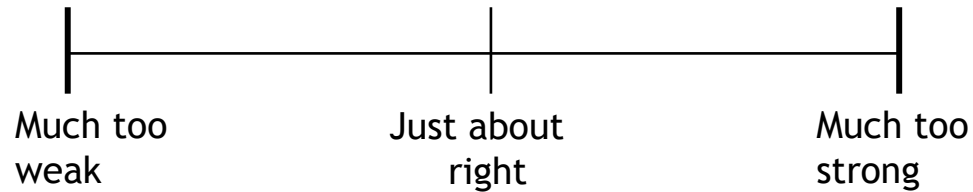
AGE: _____

-  SUPER GOOD
-  VERY GOOD
-  GOOD
-  NEITHER GOOD NOR BAD
-  BAD
-  VERY BAD
-  SUPER BAD

WHY? _____



JUST-ABOUT-RIGHT SCALE



Easy to use by assessors

Intensity rate or acceptability rate?

Just-right meaning:

Consumer panel

- Okay
- Very good
- I like the product
- Like it very much
- Highly acceptable
- Desirable
- Best for the situation
- Correct

In-house panel

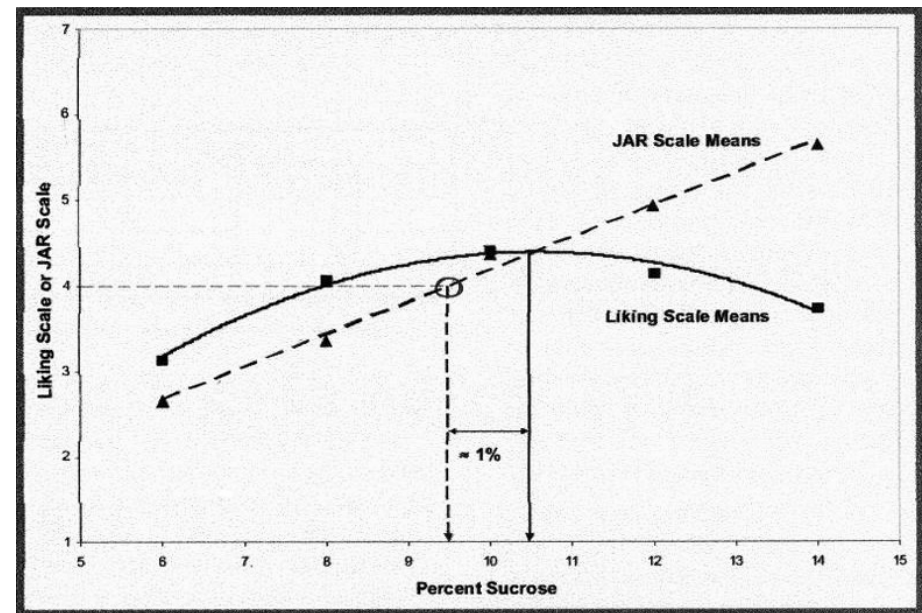
- Prefer product
- Best for the situation
- Like extremely



> JAR SCALES - BIAS EVALUATION?

Results apparently straightforward to interpret.

Will consumers rate as “just-right” those formulations that they actually like the most?



Optimal sucrose level in lemonade, Epler et al, 1998



Judgment

Health considerations (sugar, salt, fat...)

How do products differ from an ideal?

- Can an orange juice have enough fresh orange flavor?
- Are there enough chocolate chips in a chocolate cookie?

“Asking respondents to consider reasons for their preference may subsequently alter their preference.”

Wilson and Schooler, 1991



- **REPRESENTATIVE** panel selection

- Adequate panel **DIMENSION**

- **SIMPLE** is **BEAUTIFUL**

Bias results, Halo effect, Guided appreciation, Judgement

DON'T ...

- *ask a consumer for a response that he is not able to give;*

- *ask a trained panelist for a consumer response.*

“What might taste good in a laboratory setting doesn't necessarily taste good in a daily real-life setting.”

Professor Egon P. Köster, Interview Dec. 2006



Objective:

Study of the impact of changing the cork material in the degree of oxidation of a table white wine

5 alternative corks tested
18 months storage



Attribute difference-from-control test with 18

Sample	Blind assessors control Standard cork	Cork A	Cork B	Cork C	Cork D	Cork E
Mean response	+ 0,4	+0,8	- 0,5	+3,2*	+0,5	-0,2

* Significant difference 95% confidence level

6 Much weaker
0 Same
+6 Much stronger



>ATRIBUTTE DIFFERENCE TEST - APPLICATION

Objective:

Comparison of the intensity of rancid flavour in 5 different brands of potato chips subjected to temperature abuse.

Intensity rating, 18 trained assessors
15 cm line scale

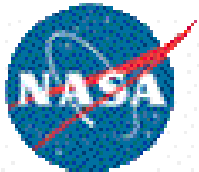
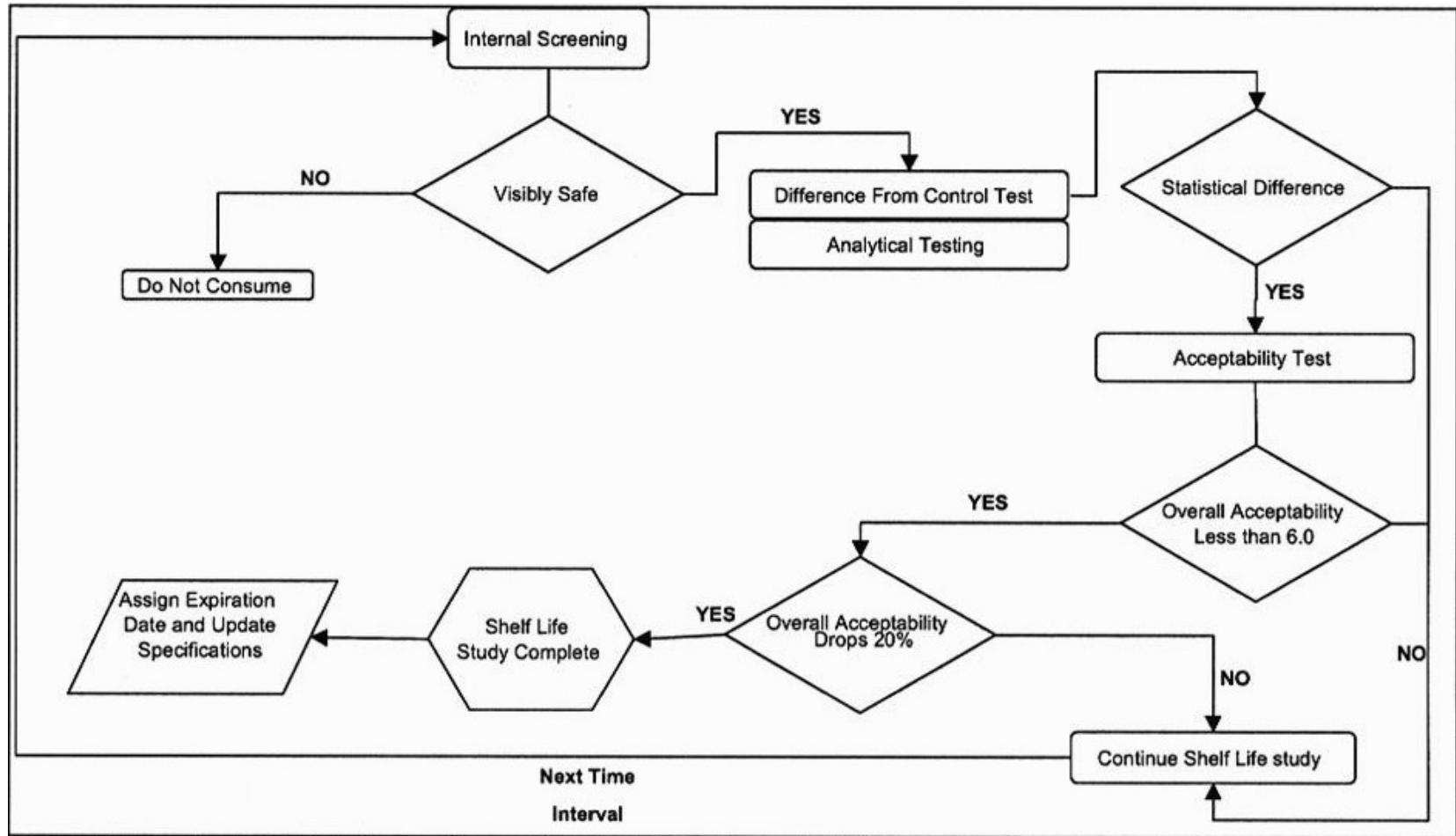


Sample	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5
Mean response	2,2a,b	6,3c	7,0c	1,5a	3.3b

Different letters - Significant difference 95% confidence level



> DISCRIMINATION TESTING - APPLICATION

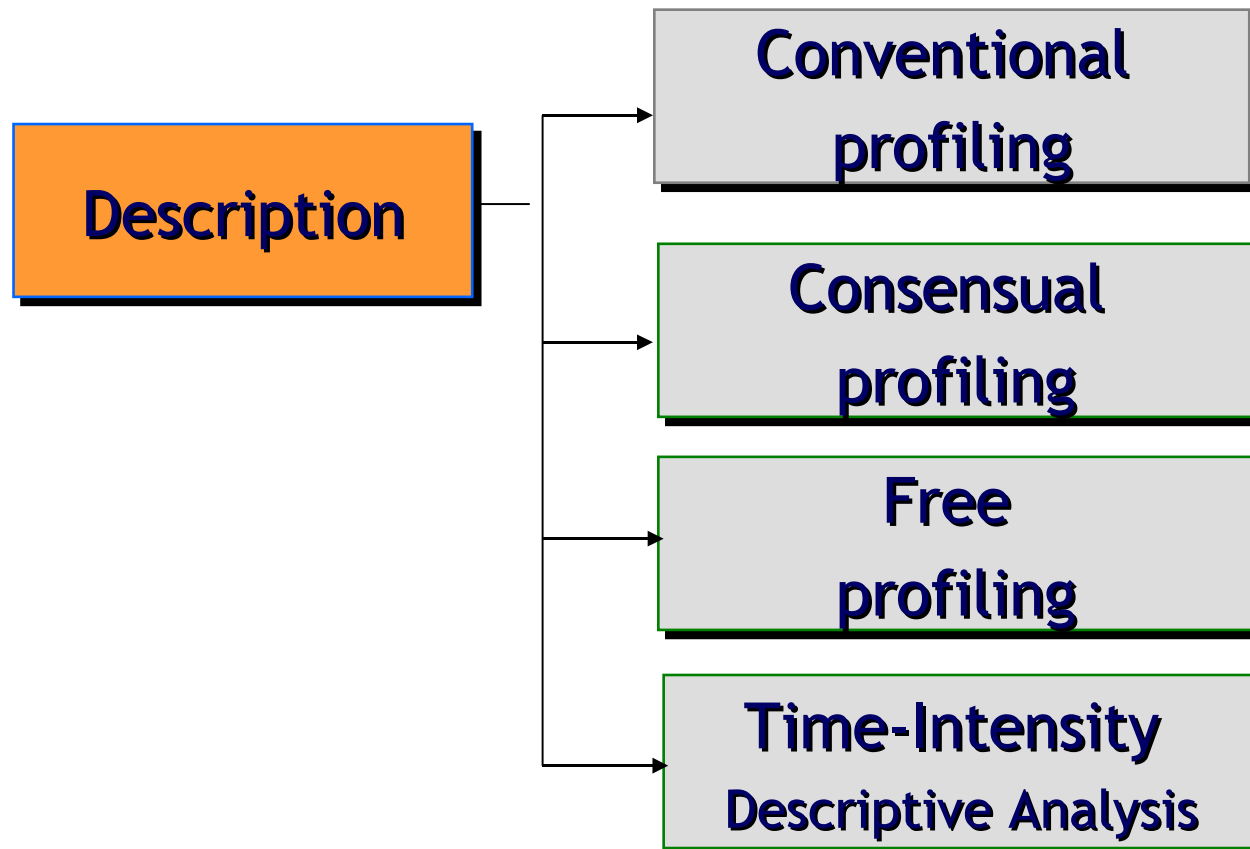


Space Food System Laboratory Shelf Life Analysis Operations Guideline



DESCRIPTIVE ANALYSIS

Attempts to provide a quantitative specification of (all) the sensory attributes of a product.



Main applications:

- Specifying sensory changes in product development as a function of ingredient, packaging or processing variables and for shelf-life and quality control evaluation.
- Data used for correlation with consumer judgment for purposes of building predictive or explanatory models of factors driving likes and dislikes.
- Data used for correlation with instrumental measures of food properties.



> CONVENTIONAL PROFILING

The most used and generally the most reliable profiling technique. Able of producing reproducible results, is suitable for research as well as for routine analysis.

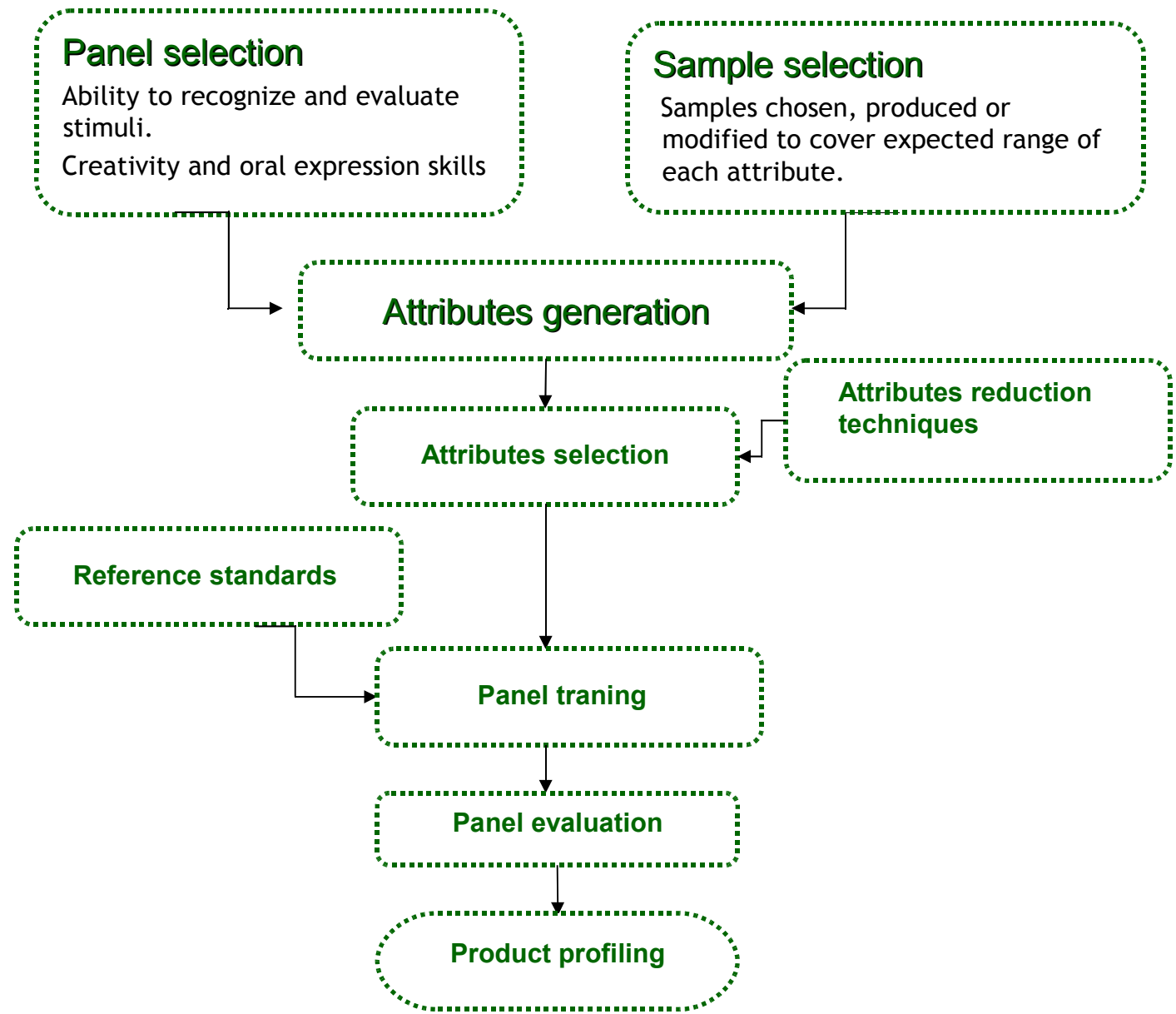
Procedures

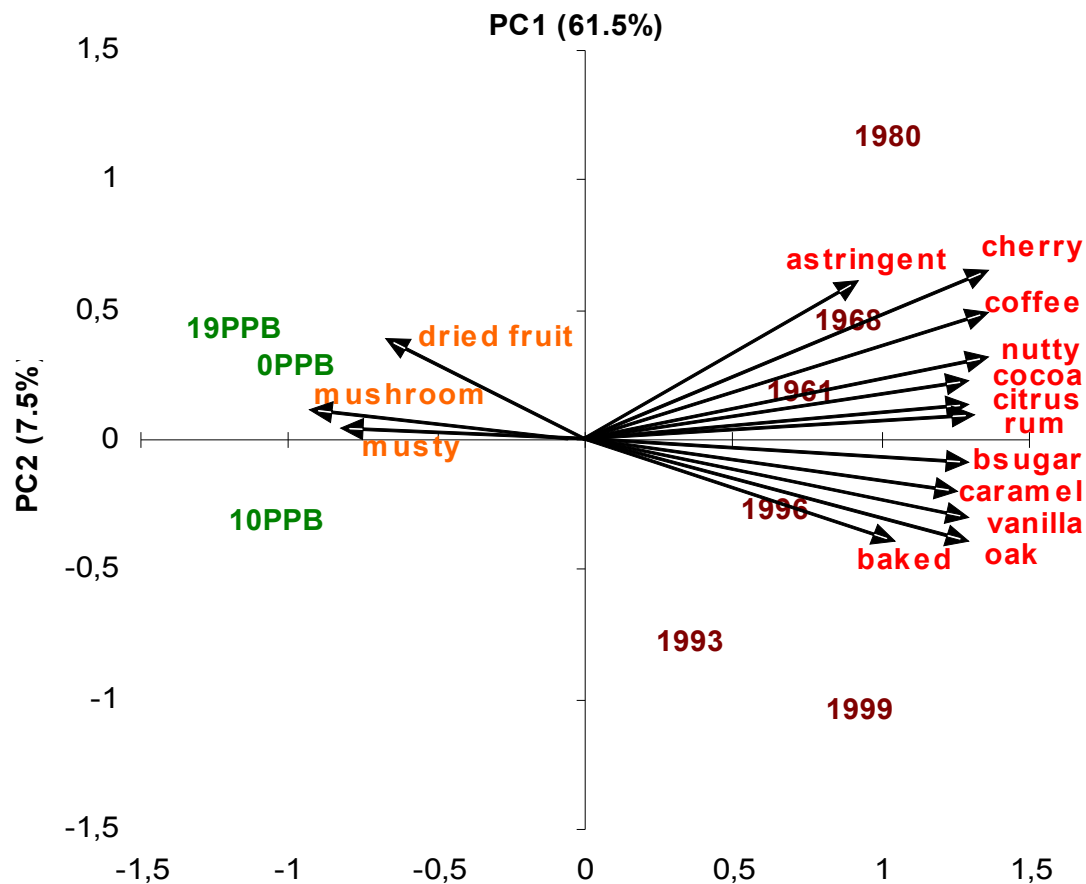
Assessors individually score each sample intensities with respect to learned absolute intensity scales

Profile is obtained by statistical treatment of obtained data (Simple graphical representation of data, ANOVA, PCA,...).



> CONVENTIONAL PROFILING





- **1980, 1968 and 1961**
cherry, nutty, coffee, astringency and cocoa flavors-by-mouth.

- **1999 and 1996**
vanilla, caramel, oak, brown sugar rum, baked and citrus

- **1993**
baked, oak and vanilla

PCs accounted for 69% of the variance

B. P. MACHADO A C. SILVA FERREIRA, HILDEGARDE HEYMANN. "Madeira wines: sensory descriptive analysis of traditional wines and wines with added sotolon". 6th Pangborn Sensory Science Symposium 7 - 11 August 2005, Harrogate, North Yorkshire, UK



CONSENSUS SENSORY PROFILING

Through consensus discussion the panel develops its own terminology and scores pertaining to the sample set presented.

Suitable for routine sensory evaluation of non-recurring products.
Many samples can be tested at relatively low cost.

FREE CHOICE PROFILING

Assessors freely choose terminology and scale to evaluate products.
Only minimum panelists training is required.

Can be used as preliminary step to develop descriptive terms, to be used in conventional profiling

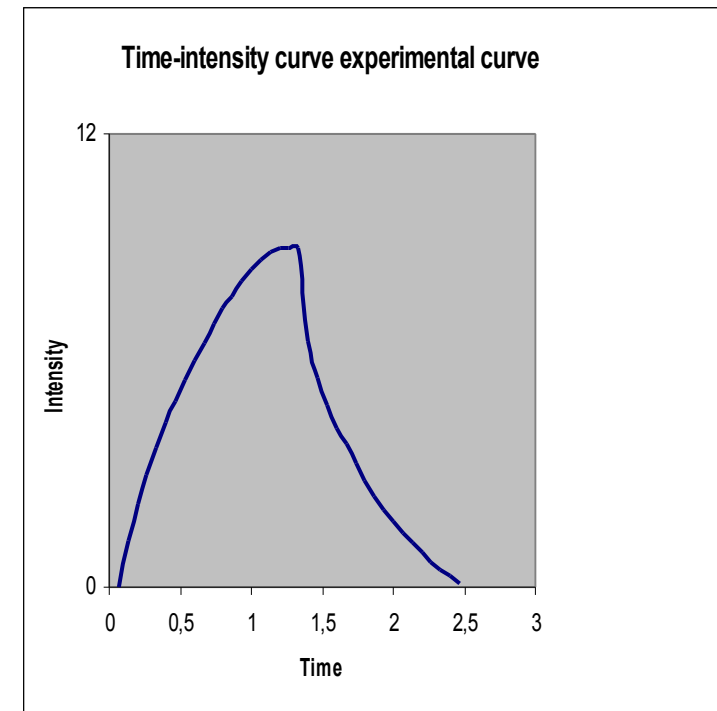


TIME-INTENSITY DESCRIPTIVE ANALYSIS

In most sensory methods, time as a dimension is not considered. However for certain attributes, the perception's intensity varies with time over a longer or shorter period.

Examples: - sweetness of sweeteners
- bitterness of beer or coffee
- astringency of wine

Protocols of evaluation must be carefully defined and requires Well-trained assessors



PREFERENCE MAPPING

Evaluate pleasure & Explore perception

- Consumers are the judges of the product, but they have neither the sensorial capacities nor the vocabulary required to express their judgments in a reliable and precise way.
- Trained sensorial experts are able to specify the nature and intensity of sensations without attaching any hedonistic value
- By usage of statistical tools the two sets of data can be linked



The best is not always the first choice

(...) As E.P. Köster sees it, consumer research comes up short when it takes note of the test ratings alone and brings the product to market readiness solely because it has the highest test results. “We often observed that such products were at first successful but then suddenly flopped. One reason for this is that the consumers’ preferences are continuously changing. Among other things people are always looking for new experiences to allay their boredom by bringing more variety into their lives. The best long term success is achieved with products that are rated positively on first consumption, but that are also complex and multi-layered. This allows the consumer to continually discover new aspects. Such products can satisfy the basic human need for excitement and stimulation.”

Professor Egon P. Köster, Interview Dec 2006



*Thank you very much for your
attention*

