

**Lesson 7: Descriptive Analysis**

**Topic 7.1: Applications of Descriptive Analysis**

**Topic 7.2: History of Descriptive Analysis**

**Topic 7.3: Principles of Descriptive Analysis**

Descriptive Language  
Abstraction and generalization in concept formation:  
Intensity Scale  
Integrated Measures

**Topic 7.4: Judge Selection**

Sensory Acuity  
Intellect  
Motivation and Availability

**Topic 7.5: Term Generation, Definitions, Evaluation Protocols and References**

Term Generation  
Definitions  
Evaluation Protocols  
References

**Topic 7.6: Scale Design**

**Topic 7.7: Panel Training**

Group Training  
Individual Training

**Topic 7.8: Criteria and Assessment of Judge Performance**

**Topic 7.9: Actual Descriptive Analysis**

Individual or Group Ratings?  
Monadic or multi-sample Presentation?

**Topic 7.10: Data Analysis**

**Topic 7.11: Final Considerations**

References

## Lesson 7: Descriptive Analysis

### Topic 7.1: Applications of Descriptive Analysis

Descriptive analysis methods involve the detection (discrimination), the description of the sensory attributes in a product (qualitative aspect), and the scaling of the intensity of these attributes (quantitative aspect) by a trained panel of 5 to 20 judges.

#### Lesson Objectives

In this lesson, we examine the principles and applications of descriptive analysis - the most comprehensive and widely used set of methods in sensory evaluation.

#### Objectives:

1. Describe the applications of descriptive analysis.
2. Examine the language and scaling components of descriptive analysis.
3. Follow the descriptive analysis process from terminology development to data analysis and reporting.
4. Explore ways to monitor judge performance.
5. Review statistical methods of descriptive data analysis.



We use descriptive analysis methods to obtain detailed descriptions of the appearance, aroma, flavor, and oral texture of foods and beverages, the skin-feel of personal care products, the hand-feel of fabrics and paper products, and more generally, the sensory properties of any product.

Some professionals extend further this broad definition to the quantitative evaluation of the 'attributes' of any product or service, but we will stick to sensory attributes in this lesson.

These pictures show a judge conducting a descriptive analysis of a skincare cream, and another evaluating a makeup sample. For these types of products, the evaluation protocol is very important because it can clearly affect how consumers perceive the product.



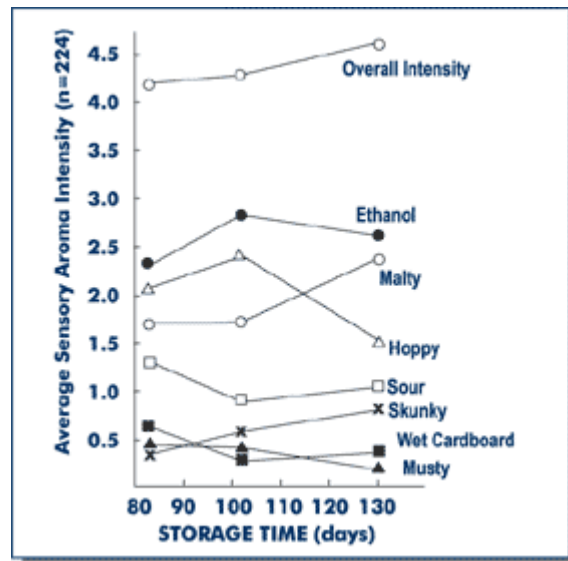
## Lesson 7: Descriptive Analysis

### Topic 7.1: Applications of Descriptive Analysis

#### Applications of descriptive analysis methods include:

- Developing an objective descriptive terminology for a product category, for which such a terminology is not currently available.
- Documenting the effects of ingredients/materials and processing variables on the sensory properties of a product.
- Defining the sensory properties of a target product and documenting the sensory properties of prototypes for new product development.
- Defining the characteristics/specifications of a control for quality assurance and quality control applications.
- Tracking a product's sensory changes over time to understand shelf-life properties.

The figure below shows how selected sensory attributes in beer vary as a function of storage time.

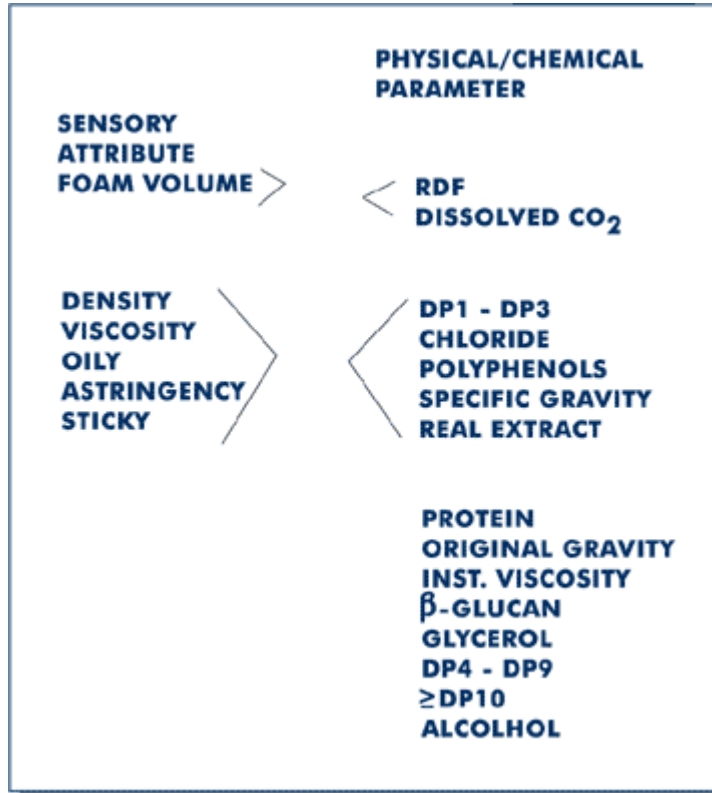


- Mapping a product's sensory attributes to relate them to instrumental, physical or chemical measurements of sensory properties. This is particularly relevant for the assessment of color, texture, and flavor in foods and beverages.

**Lesson 7: Descriptive Analysis**

**Topic 7.1: Applications of Descriptive Analysis**

The list below shows which chemical or physical property of beer relates to which mouth-feel sensory attribute in beer. For example, the real degree of fermentation (RDF) and dissolved carbon dioxide are predictors of foam volume as perceived in the mouth.

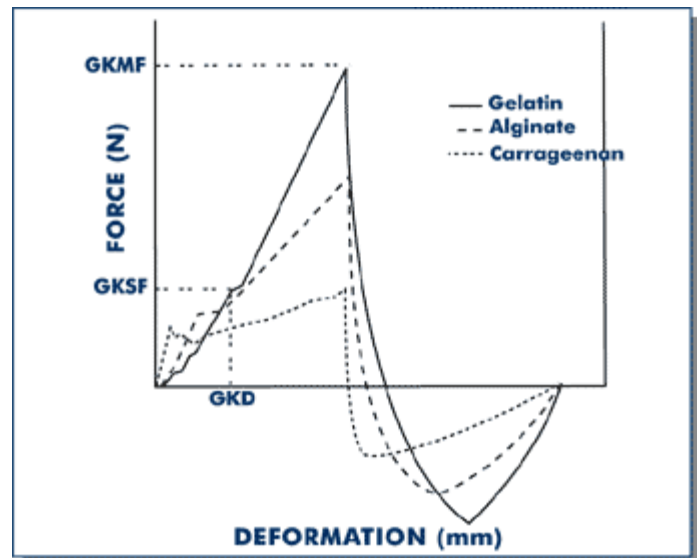


## Sensory Evaluation Methods

The next photo shows how an INSTRON can be used to measure the textural properties of gels. The instrument records the amount of force required to deform the sample and the resulting profile of force vs. deformation/time is shown next to it.



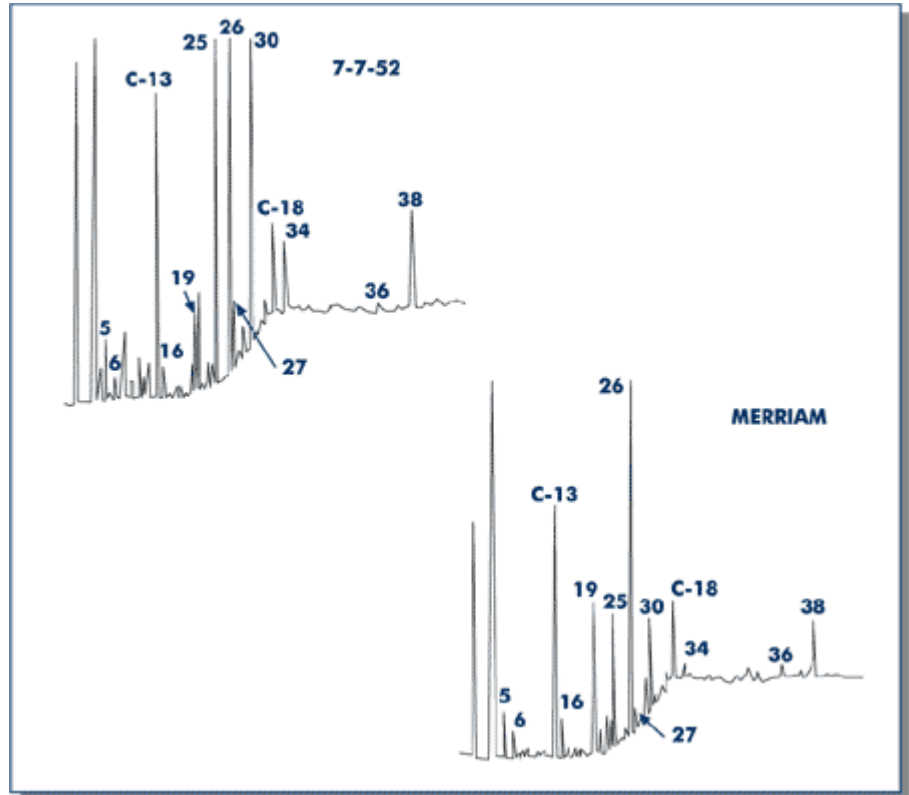
We can see that the instrument can adequately predict textural differences among gels made with gelatin, alginate or carrageenan.



## Lesson 7: Descriptive Analysis

### Topic 7.1: Applications of Descriptive Analysis

The next picture shows a gas chromatograph profile, with peaks that represent aroma compounds from the sample's headspace. We can attempt to correlate peak height or area with flavor notes perceived by a descriptive panel.



We can also pinpoint the contribution of a given compound/peak using the GC-sniff technique, whereby a non-destructive detector is used and a trained judge sniffs the compounds eluting off the GC column and makes qualitative and/or quantitative judgments.

- Document a product's attributes before a consumer test to help in questionnaire design and results' interpretation.

Quantify a product's sensory attributes to use in external preference mapping (Course 3).

## Lesson 7: Descriptive Analysis

### Topic 7.2: History of Descriptive Analysis

The first formal, systematic descriptive procedure to be developed was the **Flavor Profile Method**, at A. D. Little, as described by Cairncross and Sjostrom in 1949. These investigators showed that it was possible to select and train individuals to describe their perception of a product in some agreed sequence, leading to actionable results (in this way avoid the need for an 'expert').

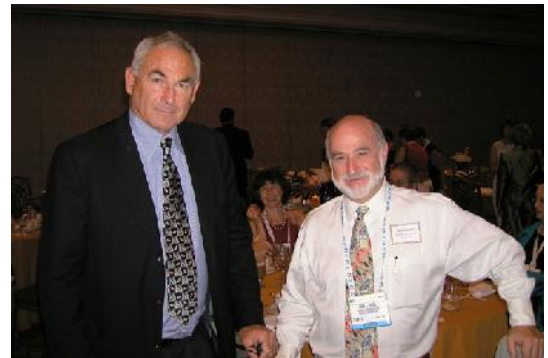
Less sophisticated, informal descriptive methods existed long before:

- Early chemists often used their senses to characterize the sensory properties of chemicals, and Gas Chromatography analysts still do.
- Expert perfumers and flavorists are longstanding examples of people using descriptors to characterize the substances they work with.
- Experts in the wine, tea, coffee, spirits and other traditional industries have long used a descriptive language (not always objective) to characterize their products.
- As consumers, we engage in informal descriptive analysis every day.

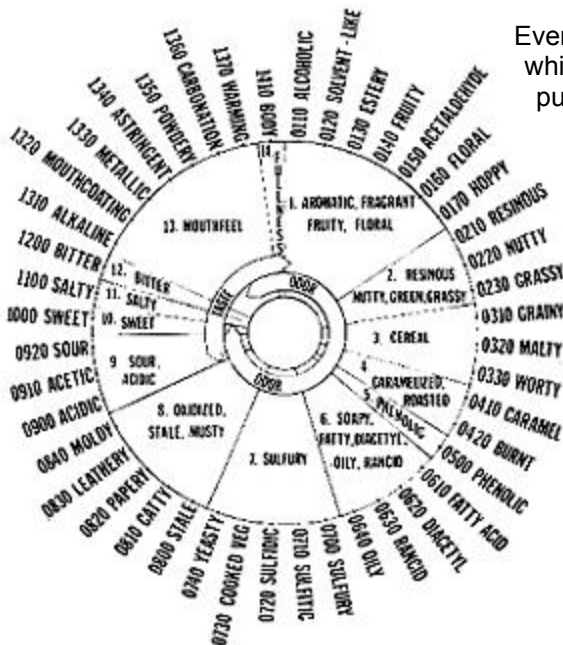
#### Other milestones in descriptive analysis were:

- The development of the **Quantitative Descriptive Analysis (QDA)** Method at Tragon, by H. Stone, J. Sidel and collaborators in 1974.

This photo shows the founders of the Tragon Corporation and creators of the QDA Method - Joel Sidel (right) and Herb Stone (left).



- The development of the Beer Flavor Wheel by American and British brewing and sensory scientists in the 1970s.



Even though the first flavor wheel was developed by the British whisky industry to describe whisky flavors, the first widely-publicized flavor wheel was the Beer Flavor Wheel shown here.

## Lesson 7: Descriptive Analysis

### Topic 7.2: History of Descriptive Analysis



The move from expert ratings of quality to descriptive analysis for wine (Maynard Amerine, UC Davis) occurred in the 1960s and 1970s and with the development of the Wine Aroma Wheel (Ann Noble, UC Davis and the California wine industry) in the 1980s.

Influential members of the Department of Viticulture and Enology at UC Davis (including Maynard Amerine - top, right) are shown tasting wine in the 1960s.

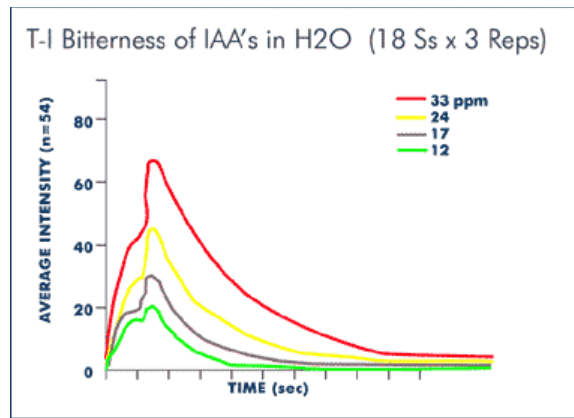
[Be sure to View the Wine Aroma Wheel on the Topic Outline](#)



- The development of the time-intensity (TI) methodology to document the intensity AND the time course of sensory attributes in the 1970s and 1980s.

The photo shows how a judge records the intensity of a sensation over time by moving a joystick along a scale as s/he tastes the sample. The other photo shows the time-intensity profiles of different concentrations of iso-alpha-acids - the bitter principle in beer.

- The development of the Spectrum Method by Gail Civille and collaborators in the late 1980s (Meilgaard, et al., 1991). The Spectrum Method, with its deliberate guidance of the panel, intensive training, absolute ratings and extensive use of references and calibration points represented a 180-degree philosophical shift from the QDA.
- The development of Free-Choice Profiling (FCP) by Williams and Langron in the early 1980s: where with this method, judges develop their own scorecard (set of terms) and use it in the descriptive analysis.





## Lesson 7: Descriptive Analysis

### Topic 7.3: Principles of Descriptive Analysis

The basic principles or components of all descriptive analysis methods are:

1. Descriptive language = qualitative aspect
2. Intensity scaling = quantitative aspect

Some descriptive analysis methods also:

- Consider the order of appearance of the attributes (e.g., the Texture Profile Method)
- Include some integrated or overall measure (of intensity, 'complexity', 'balance', 'quality', NOT liking)

The steps to most descriptive analysis methods include:

1. Term generation
2. Development of definitions and references for the attributes
3. Scorecard development (sequence of attributes, evaluation protocols, and scale)
4. Training (group and individual)
5. Assessment of judge performance and readiness
6. Descriptive analysis
7. Data analysis and reporting

## Lesson 7: Descriptive Analysis

### Topic 7.3: Principles of Descriptive Analysis

#### Descriptive Language

The perceived sensory attributes in the products are identified and described by various terms referred to as 'characteristics,' 'descriptive terms,' 'descriptors,' 'attributes,' 'notes,' or 'descriptive terminology.' Descriptive terms should be:

- Objective (not subjective)
- Unique
- Understandable
- Translatable into other languages
- References can be prepared for them

Specialists developed the flavor or aroma wheels for the whisky industry (UK), the brewing industry (UK and US) and the wine industry (US). We displayed the beer flavor wheel and the wine aroma wheel in the history segment of this lesson. We organize the terms by level of specificity, from the sensory modality stimulated (e.g., taste, smell) to general flavor categories (e.g., floral, fruity, animal, etc.) to specific flavors (e.g., apricot, pear, peach, etc.).

The panel chooses terms that best describe the differences among the products, with the right level of specificity. In some cases, fruity may be adequate. But in others, more specific descriptors of the type of fruitiness may be indicated (e.g., strawberry, apricot, orange), and those descriptors might have to be specified even further (e.g., fresh strawberries, apricot jam, orange juice, etc.). The panel may also choose to combine descriptors to cover a broader range of attributes in the products (e.g., green bean/asparagus to describe one of the vegetal characters of Cabernet Sauvignon wines).

It is critical that all judges on the panel understand the descriptive terms in the same way (**concept alignment**). We achieve this by:

- Carefully defining each term and its evaluation protocol
- Preparing references/standards for most attributes

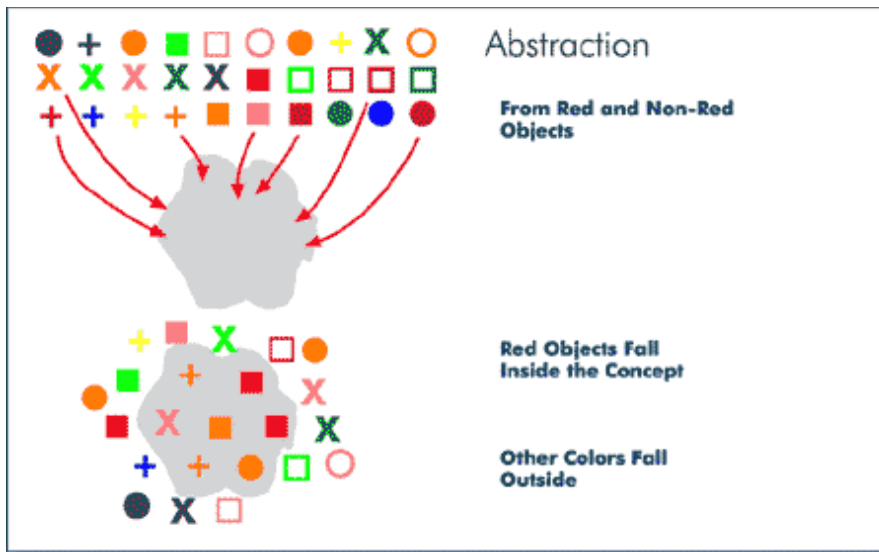
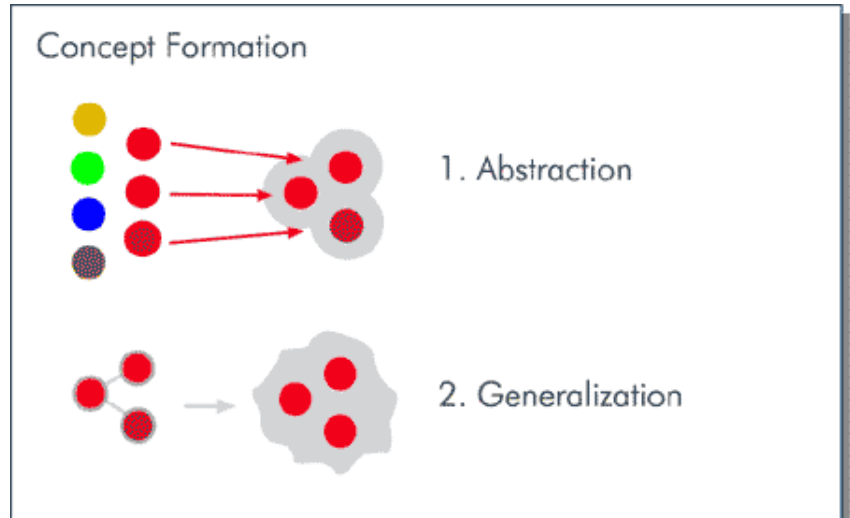
But to best be able to understand the notion of concept alignment across judges in a panel, it is helpful to examine the process psychologists refer to as 'concept formation.'

**Lesson 7: Descriptive Analysis**

**Topic 7.3: Principles of Descriptive Analysis**

**Abstraction and generalization in concept formation:**

The following pictures illustrate how the brain understands new concepts. Let's use the color RED as an example of a sensory attribute we want to explain to an individual who has no concept of what the color RED is. The brain may formulate this new concept by looking at one example of the color RED and generalizing from it. The tracing of boundaries around that example is arbitrary, though. This step is called **generalization**.



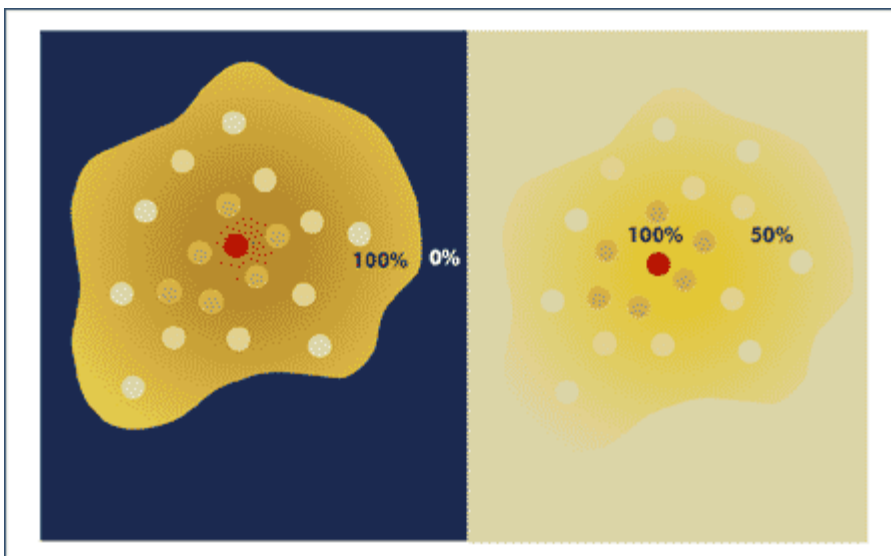
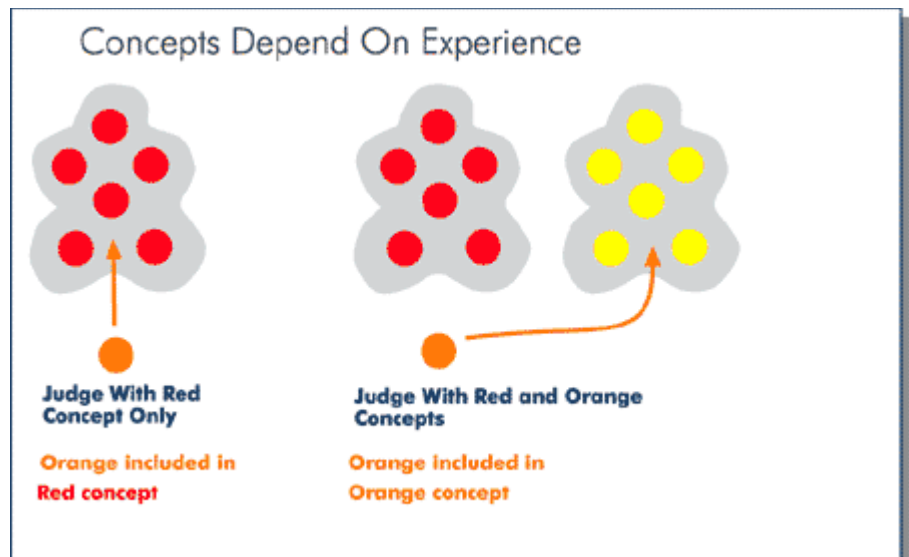
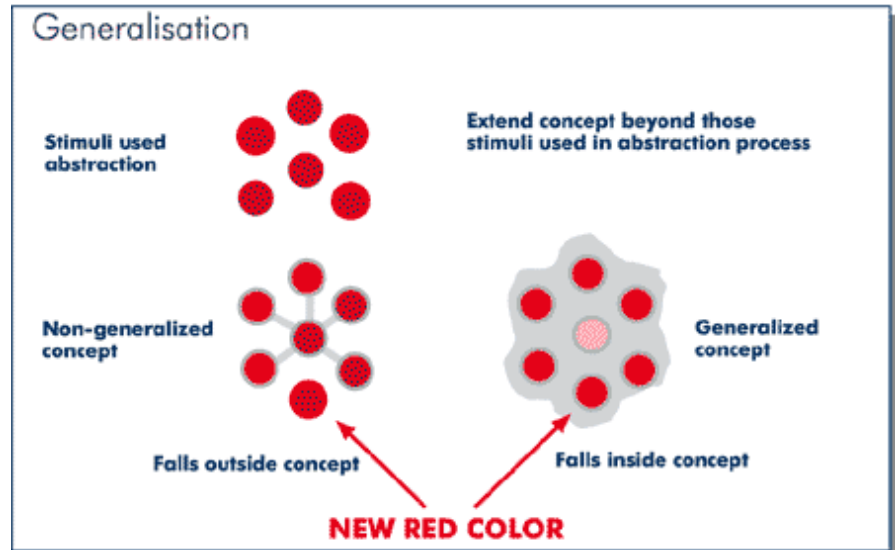
If we also present colors that are part of the RED concept and colors that are not, the brain can pinpoint the RED concept further by **abstraction** (sometimes called **discrimination** in other disciplines). The best way we achieve **concept alignment** (the same understanding of a sensory concept by all the judges on the panel) is through this two-step process of abstraction and generalization.

## Sensory Evaluation Methods

The ability to understand a concept depends on experience. Someone who is very familiar with the world of colors will understand the RED concept much easier and faster than someone who has never seen colors before. Thus, in most descriptive analysis applications, it makes more sense to work with judges who have a high degree of familiarity with the product under study than with judges who are completely naïve about it.

Some sensory concepts are more discrete than others. By this we mean that the line between two attributes (e.g., salty and sweet) can be clear cut, or it can be somewhat blurred (e.g., between peach and apricot), so that the boundaries around a given concept may be more or less difficult to set in the concept formation process.

[See Reading Assignment 7.1 on the Topic Outline](#)



## Lesson 7: Descriptive Analysis

### Topic 7.3: Principles of Descriptive Analysis

#### Who should develop the descriptive language?

The temptation is often strong in descriptive analysis to impose a set of terms/attributes to the panel and have it evaluate their intensities. But it is a risky practice at best. Product variations resulting from ingredients or processing changes rarely affect just one attribute, thus attesting to the complexity of the descriptive process. Language itself evolves. Different terms may have different meanings to different individuals. This is particularly relevant when it comes to 'technical' terms. An expert on the product may be highly comfortable with the terms, but a consumer newly recruited for a descriptive panel will not know and/or understand them as readily (e.g., a 'light-struck' flavor in beer might be better described as a 'skunky' smell for the non-expert). But allowing the panel to develop the terminology to be used in the final descriptive analysis should never be at the expense of rigorous analytical work, and the requirements for 'good' descriptors listed above always apply and should be considered carefully. If a panel agrees that terms, such as 'elegant' or 'sophisticated' be used for a wine descriptive analysis for example, it is best for the experimenter or panel leader to step in and discourage such practices (as long as s/he explains why to the panel): such terms would not likely be understood in the same way by all the judges on the panel, and they would prove hard to define.

#### Intensity Scale

Various options for scaling include:

- Line scale (10-inch)
- Numerical scale (0-10 or 0-15)
- Labeled scale (low-high, none-extreme)
- Anchor points

Some methods allow judges to use the scale any way they want (provided they discriminate among the samples AND are consistent with the rest of the panel). Other methods require that judges use the scale exactly in the same way (calibration) and extensive training is required so that judges assign approximately the same score to a given sample. The latter can be done (and may be needed for some applications), but it obviously requires a lot of training.

## Lesson 7: Descriptive Analysis

### Topic 7.3: Principles of Descriptive Analysis

#### Integrated Measures

Should integrated measures like 'overall intensity', 'degree of difference from a standard', 'complexity', 'balance', 'quality' be included in a descriptive analysis scorecard? This is a somewhat controversial issue because these parameters do not meet the requirements we set for the descriptors noted above. Except for the first two (e.g., 'overall intensity of flavor or aroma', and 'degree of difference from a standard'), they may not be objective and unique terms, and they may be difficult to define (so that the panel will rate them in the same way). Yet, consumers usually expect some measure of complexity, balance or quality by 'experts' to be available. Many of us buy wine at the supermarket or liquor store based on the ratings they received on a 100-point scale from well-known wine publications.

We should agree though that hedonic ratings (e.g., degree of liking) should be excluded from a descriptive scorecard. Because the judges on a descriptive analysis panel are not representing the consumer population, their hedonic ratings can mislead product developers or quality assurance specialists. Hedonic ratings should be obtained from consumers, not trained judges.

The following article describes a study of Lager beers, in which we had trained experts conduct a descriptive analysis of the sensory properties of the beers, as well as provide quality ratings. A high-quality beer was defined jointly between the experts and the experimenters as:

- Free of appearance and flavor defects
- True to type (e.g., meeting the expectations for the category or style)
- Well balanced, with some degree of complexity

Even with those definitions and extensive training, the experts differed somewhat in their interpretation (and ratings) of quality. So beware that quality ratings are a challenging exercise. For most consumer products, though, we feel quality ratings should be developed to guide consumers in their purchases. Consumers Union, the publishers of the Consumer Reports magazine, have had a longstanding tradition of excellence in the development of quality ratings that usually include sensory criteria. They have been successful at having experts provide accurate, reliable, and consistent evaluation of all kinds of consumer products, from running shoes to wine to handheld PDAs.

[See Reading Assignment 7.2 on the Topic Outline](#)

### Lesson 7: Descriptive Analysis

#### Topic 7.4: Judge Selection

Descriptive analysis involves relatively few judges (from 5 to 20, usually about 10), and to have confidence in the reliability and validity of the results, these judges should be screened, and then they must be trained. The purpose of screening potential judges should be to ensure that they can perceive differences at better than chance level among products of the kind that is to be evaluated by descriptive analysis. Furthermore, the experimenter should ensure that motivation, sound intellect, and availability complete the potential judge's profile.

#### Sensory Acuity

Performance in sensory tests with simple stimuli (basic tastes, smells) does not adequately predict performance with actual products; therefore, such tests are generally a waste of the investigator's time and resources. Instead, judges should be screened with actual products. Difference tests with samples that are significantly, but barely, different are ideal. In the QDA method, for example, 20 to 30 discrimination trials fielded over 2 or 3 days are recommended to select judges from a pool of individuals naive to sensory testing. Expect that about 30% of those who volunteer will fail to meet the chance probability requirement. Scaling exercises may also be useful on what is believed to be key sensory attributes and representative samples for the product under study. Keep in mind that screening is intended primarily to eliminate non-discriminators and secondarily to familiarize the judges with the sensory properties of the product. **It does not eliminate the need for training the panel.**

Note that most descriptive analysis methods do not call for screening potential judges for sensory acuity. But it just makes sense... so to speak.

#### Intellect

A comment about the type of individual to look for in descriptive analysis: in addition to having good sensory abilities, ideal judges are those with good intellectual abilities. This is because descriptive analysis is a sensory as well as an intellectual exercise that requires a great deal of analytical skill and concentration.

#### Motivation and Availability

In the end, motivation and availability might be the most important features to look for in your selection of judges. A motivated and available judge will actively participate in the term generation process and in group discussions and training, follow instructions carefully, be on time for his/her sessions and will not have to make up any missed sessions. As always, best practices in sensory evaluation begin with good common sense. Maintaining panel motivation throughout the descriptive analysis process by providing feedback to the judges is also critical.

## Lesson 7: Descriptive Analysis

### Topic 7.5: Term Generation, Definitions, Evaluation Protocols and References

Descriptive analysis methods vary widely with regard to term generation, definitions, evaluation protocols, and use of references. Some methods deliver the list of attributes to the judges, along with corresponding definitions, evaluation protocols and (multiple) references, and the early stages of the descriptive analysis are spent training the panel on those. Other methods rely entirely on the panel for term generation, definitions, evaluation protocols, and may or may not advocate the use of references. Again common sense should prevail here, and we will argue that the road to travel may be somewhere between those two extremes.



#### Term Generation

Unless the method stipulates that the experimenters pre-select the attributes to be evaluated and merely teach the panel how to rate them, it is up to the panel to come up with a sensory lexicon describing the sensory attributes of the product. This is typically done by showing the panel a set of samples representative (in terms of both range and intensities of attributes) of the samples to be tested in the actual descriptive analysis over several sessions. With each sample, judges are encouraged to propose terms describing the sensory attributes, usually with a focus on those attributes that tend to differ among samples.

Both photos show descriptive panels at work during the term generation phase.



## **Lesson 7: Descriptive Analysis**

### **Topic 7.5: Term Generation, Definitions, Evaluation Protocols and References**

#### **Definitions**

All the terms in the scorecard should be carefully defined. In some cases, the attribute may be straightforward (e.g., salty, hard, shiny). But in most cases, the attribute may be interpreted in a number of ways, and this is why it is important to define it as precisely as possible. For example, if a panel chooses to include the term 'strawberry aroma' in the scorecard, the panel must define what type of strawberry aroma it is referring to: for example, is it fresh strawberry, overripe strawberry, or strawberry jam? Otherwise, the risk of poor concept alignment may be significant.

#### **Evaluation Protocols**

The same logic applies to the careful development of evaluation protocols. To ensure that everyone on the panel is working as one instrument, it is important that all judges evaluate each attribute in the same way. For example, an evaluation of chewiness should specify how to chew the sample (e.g., Using molar teeth? How many chews? Etc.).

## Lesson 7: Descriptive Analysis

### Topic 7.5: Term Generation, Definitions, Evaluation Protocols and References

#### References

Aroma and flavor by mouth terms selected for descriptive analysis and composition of the corresponding reference standards.	
Term*	Composition of reference standard*
1. Fresh berry (strawberry, raspberry, black currant)	5 mL reb berry fruit drink (Capri Sun™) + 3 mL black currant syrup (Vedrenne™)
2. Berry jam (strawberry, raspberry, blackberry)	6.5g each of strawberry, raspberry and blackberry jam (Empress™)
3. Cherry	5 mL cherry drink (Hi-C™)
4. Prune	10 mL prune juice (Town House™)
5. Spicy (black pepper, cloves)	pinch of black pepper + 2 cloves
6. Mint / eucalyptus	2 cm <sup>2</sup> green mint + 4 cm <sup>2</sup> eucalyptus leaf
7. Earthy (potato, mushroom)	7.5 mL canned potato liquor + 10 mL canned mushroom liquor (Town House™)
8. Leather	4 cm <sup>2</sup> leather
9. Vegetal (green bean, green tea)	10 mL canned green bean liquor (Town House™) + 1.2 green tea (Dynasty™)
10. Smoke / Tar	0.01 mL liquid hickory smoke (Wright's™) + 1 g tar
11. Berry by mouth	1 mL I.F.F.™ strawberry extract in 150 mL Pinot noir
12. Bitterness	0.2 g caffeine in 150 mL water/0.4 g caffeine in 150 mL Pinot noir
13. Astringency	0.6 g aluminum sulfate in 150 mL water/1.2 g aluminum sulfate in 150 mL Pinot noir
In 30 mL Mountain Castle Burgandy™ (unless otherwise specified)	

Some descriptive methods call for the use of references. We believe that they are particularly indicated for appearance and flavor attributes (color, aroma, taste, chemical irritation) because they can readily be developed for those modalities. For other types of attributes, such as texture and mouth-feel, it may be more challenging to develop a reference to illustrate an attribute.

The table to the left shows a list of terms used in the descriptive analysis of Pinot noir wines and the recipes for making the corresponding references. The photo shows members of a descriptive panel sniffing wine aroma references. Those were made by spiking a neutral base wine with specific items that illustrate the flavors in the scorecard (for example, some freshly cut grass soaked in base wine as a reference for 'grassy' flavor, or a few drops of vanilla extract in base wine to represent 'vanilla' flavor, etc.).

Judges are asked to carefully evaluate each reference for accuracy (does it really pinpoint the flavor of interest?) and to memorize it. This way, the experimenter may achieve concept alignment across the panel.

Returning to our discussion of concept formation above, we can extend the principles of concept formation to the use of references in descriptive analysis. In most cases, we resort to generalization to teach a sensory concept to the judge. This means that we make a reference that illustrates the flavor note of interest (and the judge must generalize that concept from the reference). For example, we might take a neutral sample and spike it with an item or a chemical to impart the flavor note of interest. In other instances, however, we may not be able to manufacture a sample that illustrates the attribute (for most texture attributes, for example), and we typically resort to abstraction to illustrate a concept. We show the panel samples with the attribute and samples without it, and hope that the judge can understand the concept by abstraction

See Reading Assignment 6.1 from Lesson 6, and Reading Assignment 7.3 on the Topic Outline for more details on the use of references in descriptive analysis.



## Lesson 7: Descriptive Analysis

### Topic 7.6: Scale Design

Most category scales are well suited for descriptive analysis. They may be graphic, line scales or numerical scales. They are typically anchored with terms such as "low" and "high," or "none" and "extreme," and may include additional terms along the scale.

Our recommendations are:

1. Make sure that the scale has enough categories to adequately discriminate among the samples.
2. **KEEP IT SIMPLE.** As shown in our scaling lesson, many features on a scale are not useful and may actually confuse the judge.

Even more important than the design of the scale is to know how the judges will use it. Depending on the descriptive method, the expectations of the panel will range from relative to absolute ratings, the latter requiring extensive calibration with multiple references.

## **Lesson 7: Descriptive Analysis**

### **Topic 7.7: Panel Training**

A number of textbooks and practitioners consider that the steps we have covered so far ARE the training of the panel. Big mistake! Discussing the features of the products, agreeing on some references for the attributes, and designing testing protocols and evaluation scales by no means trains a panel to perform a descriptive analysis. No wonder so many (published) descriptive data sets are blatant examples of poor judge performance (see criteria of judge performance in the next section)...

The steps covered so far have allowed us to develop a set of tools to conduct a descriptive analysis. The panel now needs to be trained in the use of these tools. For training a descriptive panel, we recommend conducting first group sessions and then individual sessions.

#### **Group Training**

In the group phase of the training, we recommend presenting selected samples to the panel and having the judges discuss openly how they would rate the attributes across the samples. This is particularly helpful to resolve concept alignment issues (when an attribute may not quite yet have the same perceptual meaning for all the judges). You may at first ask the judges to rank the samples in order of increasing intensity for each of the attributes in the scorecard, and subsequently have them give ratings using the scale which was selected. By discussing their rankings and/or ratings, and going back to the samples, judges in the panel slowly but surely will achieve the needed level of concept alignment for the attributes in the scorecard.

#### **Individual Training**

In individual training sessions, the judges are asked to evaluate samples the way they will in the final descriptive analysis. This is based on the "practice makes perfect" belief. By becoming very familiar with the samples, the scorecard and the evaluation protocols, the judges simply get better at the task. Furthermore, their performance is monitored throughout (or at least towards the end of) training (using the statistical tools outlined in the next section) to ensure that they are ready to perform the final descriptive analysis. If they are not ready, they must be trained further.

## Lesson 7: Descriptive Analysis

### Topic 7.8: Criteria and Assessment of Judge Performance

We typically consider three criteria of judge performance:

1. Ability to discriminate among the samples
2. Reproducibility
3. Consistency with the rest of the panel - concept alignment

If a descriptive analysis is designed with samples, judges, replications and their two-way interactions as sources of variation, we can use the ANOVA procedure to analyze the data for each attribute, and obtain measures of the performance criteria listed above for each attribute. Here are the F-ratios we should examine for each criterion.

1. Ability to discriminate among the samples: F-ratio for the samples. If the judges adequately discriminated among the samples, the F-ratio will be significant for that attribute.
2. Reproducibility: F-ratio for the replications AND F-ratio for the judge by replication interaction. The F-ratio for the replications is an indication of whether the overall means for each replication are different or not. Obviously, this F-ratio will not be significant if the panel was reproducible across the replications. But this can be the case if some judges went up, some were consistent, and others went down, because we are only looking at the overall means (across all judges). This is why we must examine the judge by replication interaction as well. If it is significant, it tells us that some judges were not reproducible (went up or down across replications). Ideally, if the judges were reproducible, neither the replications F-ratio nor the judge by replication F-ratio should be significant.
3. Consistency with the rest of the panel - concept alignment: F-ratio for the judge by sample interaction. This interaction tells us whether the judges have the same trends across the samples for a given attribute: do they all agree that sample A was stronger than sample B, or do we have a couple of judges who scored it the other way? If that is the case, the judge by sample interaction will be significant. Good concept alignment (consistency across judges) results in a non-significant judge by sample F-ratio.

So to summarize, our judges are performing well if:

1. The samples F-ratio is significant.
2. The replications F-ratio and the judge by replication F-ratio are not significant.
3. The judge by sample F-ratio is not significant.

Note that we do not pay much attention to the judges' F-ratio. It will be significant in most instances, because judges use different parts or ranges of the scale (unless the descriptive method requires tight calibration, with judges required to give the same rating on the scale to a given sample - e.g., Spectrum Method).

**Lesson 7: Descriptive Analysis**

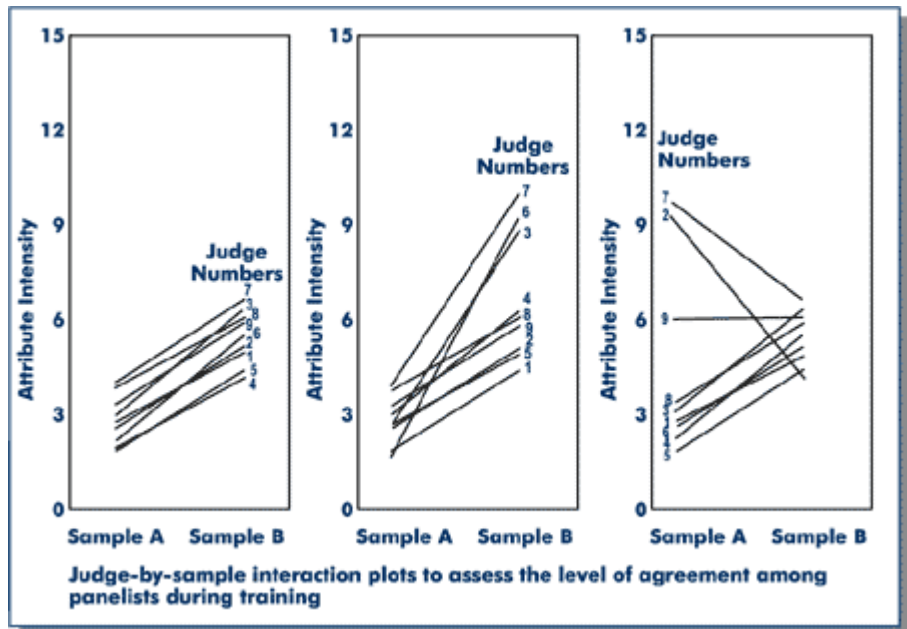
**Topic 7.8: Criteria and Assessment of Judge Performance**

If any of the performance criteria shown above are not met, the experimenter must examine individual judge's data to determine who is not performing adequately and may require additional training.

- To assess the judge's ability to be on target, we can determine how well this judge's ratings match the accepted intensity of a control or reference (assuming the descriptive method requires quantitative calibration with references). We can measure the so-called 'bias' of each judge,  $d = x - \mu$ , where  $d$  is the deviation from the target,  $x$  is the observed judge's value, and  $\mu$  is the value for the reference or target. The absolute bias is derived by removing the + or - sign from the bias:  $|d| = |x - \mu|$ . If several evaluations were conducted, the judge's own variability about his/her own rating is the standard deviation: Judges who perform well will have both low bias and low variability. By plotting the judges' statistics over time, one can identify those judges who need retraining or calibration.

$$s = \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)}$$

- If the judge by sample interaction is significant in an ANOVA procedure (for one or more of the attributes), it is necessary to examine plots of the data to determine the source(s) of the interaction. The figure below shows three hypothetical plots of judge by sample interactions. In the left plot, the interaction is not significant. All the judges rated the samples in the same direction and with more or less the same degree of intensity. In the right plot, however, three judges did not pick up on the increase in intensity between sample A and sample B, and scored them the same (judge #9) or in the opposite direction (judges #2 and 7). In this case, the interaction is significant, concept alignment is poor and those 3 judges must be retrained and calibrated. In the center plot, the slopes of the judges differ somewhat and this might lead to a significant F-ratio. This is not a major cause for concern because most descriptive methods are not absolute methods that require all judges to give the exact same rating on the scale to a given sample (except the Spectrum Method). Thus, these methods are functional with judges picking up on the trend among the samples (relative ratings). Generally, a highly significant ( $p < 0.001$ ) judge by sample interaction indicates the need for review of terminology and definitions, more frequent use of references, and more training. In those instances when significant judge by sample interactions are spotted after the fact (that is, on the actual descriptive analysis data), one must look for judges who consistently falter (for several of the problematic attributes) and can remove them from the data analysis (for all attributes). The interactions should disappear with a 'clean' set of judges. Be careful not to remove too many judges, though, otherwise the remaining pool may not provide an accurate measurement. I would not remove more than 25% of the judges from a data analysis. One may not be so lucky to find those judges, though, and be in a situation when poor training led to poor descriptive analysis... That would mean starting over from scratch, and doing it right the second time around! Note that some methods strongly advise against removing judges from the data analysis (e.g., the QDA Method).



## **Lesson 7: Descriptive Analysis**

### **Topic 7.8: Criteria and Assessment of Judge Performance**

Multivariate methods of analysis offer other, more complex alternatives for the assessment of judge performance.

Instead of conducting a principal component analysis on the mean descriptive ratings (see data analysis section below), one can apply Generalized Procrustes Analysis to the individual ratings and examine the position of each judge in relation to the other judges for the various attributes in the scorecard. Ideally, if we plot attribute X, as rated by each individual judge on the PCA biplot showing the variables (the attributes), we should see all the attribute X vectors bunched together. If the attributes X of some of the judges are found elsewhere on the plot, this would suggest poor concept alignment.

Multidimensional scaling can also be used to analyze distances among the profiles derived by the judges on the panel, and cluster analysis can be used to group judges based on the similarity of their sensory profiles across a set of samples.

[See Reading Assignment 7.4 on the Topic Outline](#)

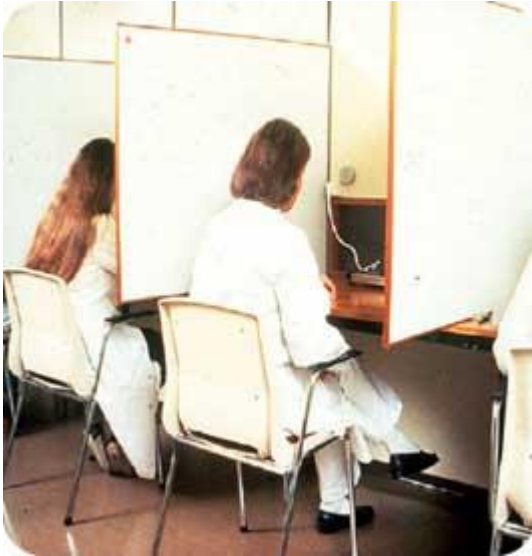
For more information on these multivariate approaches, please consult Naes and Risvik, 1996. Multivariate analysis of data in sensory science. Elsevier, Amsterdam, 348 pp.

## Lesson 7: Descriptive Analysis

### Topic 7.9: Actual Descriptive Analysis

The actual descriptive analysis follows the design selected by the experimenters or the panel, depending on the method. That includes the scorecard (sequence of attributes, evaluation protocol, and scale(s)), the number of samples and of replications, and the number of samples evaluated per session.

#### Individual or Group Ratings?



Almost all descriptive analysis methods have judges evaluate the products individually in booths. But one method - the original Flavor Profile Method - has judges rate the products as a group and give consensus ratings. Obviously, the need for individual training is not warranted in that case, but training is still extensive, among other things because it takes a lot of it for judges to use the scale in the same way and to come to a consensus easily and systematically (on each attribute, across all samples...).

Two descriptive analysis configurations are shown here, one with judges in isolated booths, and the other in a round table format. Note that a group configuration as shown on the slide to the left can sometimes be converted to a booths configuration by erecting some partitions between the judges.





## **Lesson 7: Descriptive Analysis**

### **Topic 7.9: Actual Descriptive Analysis**

#### **Monadic or multi-sample Presentation?**

An important decision must be made regarding the sample presentation mode before the actual descriptive analysis. Should the judge be given one sample at a time and rate all the attributes for that sample before receiving the next sample? Or should the judge be given all the samples at once and rate each attribute (or group of similar attributes) across the samples before moving to the next (set of) attribute(s)? We call the first option the monadic presentation mode and the second option the multi-sample presentation mode.

The following article presents the results of our research comparing these two presentation modes in a descriptive analysis of milk chocolates. The sensory profiles generated with the two presentation modes were very similar (as assessed by spider web plot and principal component plots). Time-wise, the multisample presentation mode took significantly longer to complete (e.g., 26 min vs. 21 min). Judges discriminated slightly better and were more reproducible with the multisample mode, but quality of concept alignment did not differ between the two protocols. We conclude that each method has its advantages and limitations, and that choice of the best presentation mode may depend on the descriptive method and the product. But overall, our recommendation is to stick with the monadic presentation mode for simplicity's sake.

[See Reading Assignment 7.5 on the Topic Outline](#)

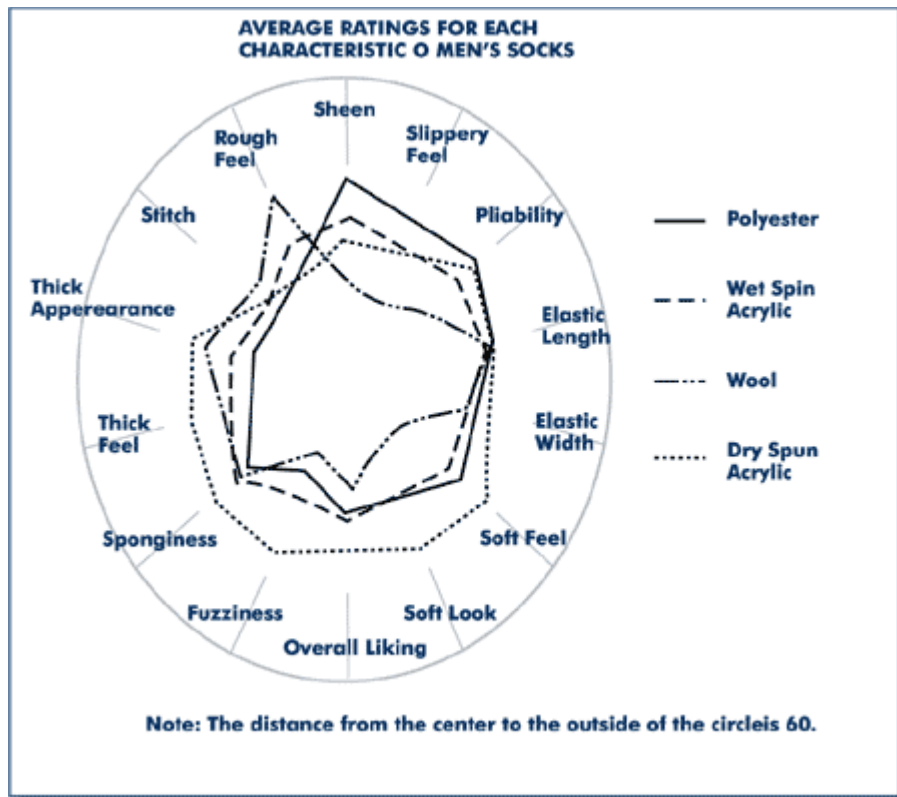
## Lesson 7: Descriptive Analysis

### Topic 7.10: Data Analysis

We have already addressed many aspects of the data analysis process with respect to the assessment of judge performance. But the main purpose of the data analysis is to determine how the sensory attributes evaluated differ among the samples. A typical data analysis sequence would go like this:

1. Multivariate analysis of variance (MANOVA) - not commonly run due to the relative sophistication of the analysis is a deterrent to many
2. Analysis of variance (ANOVA or PROC GLM) for each attribute
3. Means and multiple mean comparisons for each attribute
4. Multivariate analysis of the matrix of mean attribute ratings across the samples (PCA, cluster analysis, MDS, etc.) or of individual ratings for all the attributes across the samples (Procrustes analysis)

We use a number of tables or graphical outputs to report the results. One famous feature of the QDA Method is the Spider Web Plot that displays the mean intensity ratings of the samples. It allows for a very striking and accurate visualization of the sensory profiles of the samples and of the differences among them.



This shows a spider web plot for a descriptive analysis of men's socks.

## Lesson 7: Descriptive Analysis

### Topic 7.11: Final Considerations

How long does it take to put together and run a descriptive analysis? The answer is A LONG TIME!

All the steps we have covered are time-consuming. The speed at which we move through them depends on the panel's qualifications (as determined by screening) and motivation. Here is a typical timeframe for most descriptive analyses carried out in my laboratory for a fairly complex product, with all sensory attributes being included (e.g., appearance, flavor - taste, smell, chemical irritation - texture and mouth-feel).

- Term generation: 4 to 8 one-hour sessions.
- Definitions, evaluation protocols, references, scale: 4 to 6 one-hour sessions.
- Training: group - 4 to 6 one-hour sessions; individual - 4 to 6 fifteen- to thirty-minute sessions
- Actual descriptive analysis: this depends on the number of samples in the design, the length of the scorecard, the number of replications, the number of samples evaluated per session, etc. For example, 15 samples in the design, 20 attributes in the scorecard, 3 replications and 5 samples per session would take 9 sessions of about 20 to 30 minutes in length.
- Total for the example given: 26 to 35 sessions! If our panel met every day of the work week, the descriptive analysis would take 5 to 7 weeks. More typically our panels meet or test about 3 times a week, and a descriptive analysis takes 8 to 12 weeks.
- Don't forget to add time for screening potential judges if you did not yet have a pool to draw from, and for data analysis and reporting.

Note that the critical point comes at the end of the training phase. If we have done our job right up to that point, judge performance will be adequate and the panel is deemed ready for the actual descriptive analysis. It is often the case, however, that some of the judges are not quite up to speed at the end of the training and need additional, individual coaching. That can delay the start of the analysis by a week. But this will be the best use of a week you will ever make. The alternative is rushing it and collecting... flawed and hence useless data!

[See Reading Assignment 7.6 on the Topic Outline for an example of descriptive analysis from start to finish.](#)

## Lesson 7: Descriptive Analysis

### References

Lawless, H. T., & Heymann, H. (1998). *Sensory Evaluation of Food. Principles and Practices*. New York: Chapman & Hall.

Meilgaard, M., Civille, G. V. & Carr, B. T. (1991). *Sensory Evaluation Techniques* (2nd ed.). Boca Raton: CRC Press.

Naes T. & Risvik, E. (1996). *Multivariate analysis of data in sensory science*. Elsevier, Amsterdam.

O'Mahony, M. (1986). *Sensory Evaluation of Food: Statistical Methods and Procedures*. New York: Marcel Dekker.

Stone, H. & Sidel, J. (1993). *Sensory Evaluation Practices* (2nd ed.). San Diego, CA: Academic Press.