Chapter 10

Descriptive Analysis

Abstract This chapter describes the potential uses for descriptive analysis in sensory evaluation. We then discuss the use of language and concept formation as well as the requirements for appropriate sensory attribute terms. This is followed by a historical review of the first descriptive analysis technique, the Flavor Profile. We then describe the Texture Profile, as well as proprietary descriptive methods such as Quantitative Descriptive Analysis and the Spectrum method. We then lead the reader through a step-by-step application of consensus and ballot-trained generic descriptive analyses. We then highlight and discuss some of the studies comparing conventional descriptive analysis technique. This is followed by an in-depth discussion of the variations on the theme of descriptive analysis such as free choice profiling and flash profiling.

I want to reach that state of condensation of sensations which constitutes a picture.

—Henri Matisse

Contents

10.1 Introduction

Descriptive sensory analyses are the most sophisticated tools in the arsenal of the sensory scientist. These techniques allow the sensory scientist to obtain complete sensory descriptions of products, to identify underlying ingredient and process variables, and/or to determine which sensory attributes are important to acceptance. A generic descriptive analysis would usually have between 8 and 12 panelists that would have been trained, with the use of reference standards, to understand and agree on the meaning of the attributes used. They would usually use a quantitative scale for intensity which allows the data to be statistically analyzed. These panelists would not be asked for their hedonic responses to the products. However, as we will see in this chapter, there are several different descriptive analysis methods and, in general, these reflect very different sensory philosophies and approaches. Usually, descriptive techniques produce objective descriptions of products in terms of the perceived sensory attributes. Depending on

H.T. Lawless, H. Heymann, *Sensory Evaluation of Food*, Food Science Text Series,
DOI 10.1007/978-1-4419-6488-5_10,

the specific technique used, the description can be more or less objective, as well as qualitative or quantitative.

10.2 Uses of Descriptive Analyses

Descriptive analyses are generally useful in any situation where a detailed specification of the sensory attributes of a single product or a comparison of the sensory differences among several products is desired. These techniques are often used to monitor competitors' offerings. Descriptive analysis can indicate exactly how in the sensory dimension the competitor's product is different from yours. These techniques are ideal for shelf-life testing, especially if the judges were well trained and are consistent over time. Descriptive techniques are frequently used in product development to measure how close a new introduction is to the target or to assess suitability of prototype products. In quality assurance, descriptive techniques can be invaluable when the sensory aspects of a problem must be defined. Descriptive techniques tend to be too expensive for day-to-day quality control situations, but the methods are helpful when troubleshooting major consumer complaints. Most descriptive methods can be used to define sensory–instrumental relationships. Descriptive analysis techniques should never be used with consumers because in all descriptive methods, the panelists should be trained at the very least to be consistent and reproducible.

10.3 Language and Descriptive Analysis

There are three types of language, namely everyday language, lexical language, and scientific language. Everyday language is used in daily conversations and may vary across cultural subgroups and geographical regions. Lexical language is the language found in the dictionary and this language may be used in everyday conversations. However, few people use primarily lexical language in conversation. For most of us lexical language is best represented in our written documents. Scientific language is specifically created for scientific purposes and the terms used are usually very precisely defined. This is frequently the "jargon" associated with a specific scientific discipline.

The training phase of most descriptive analysis techniques includes an effort to teach the panel or to have the panel create their own scientific language for the product or product category of interest. Psychologists and anthropologists have argued for years about the interplay between language and perception. An extreme view is that of Benjamin Whorf (1952) who said that language both reflects and determines the way in which we perceive the world. On the other side of the coin are psychologists who say that perception is largely determined by the information and structure offered by stimulation from the environment. Words serve merely as instruments to convey our perceptions to other people. There is evidence that humans learn to organize patterns of correlated sensory characteristics to form categories and concepts. The concepts formed are labeled (given language descriptions) to facilitate communication.

Concept formation is dependent on prior experience. Thus different people or cultures may form different concepts from the same characteristics. Concepts are formed by a process involving both abstraction and generalization (Muñoz and Civille, 1998). A number of studies have shown that concept formation may require exposure to many similar products, certainly if the end result is a desire to align a concept among a group of people (Ishii and O'Mahony, 1991). A single example may define the prototype for the concept (usually called a descriptor in sensory studies) but does not necessarily allow the panelists to generalize, abstract, or learn where the concept boundaries are. To generalize and learn to distinguish weakly structured concepts (such as creaminess) the panelists should be exposed to multiple reference standards (Homa and Cultice, 1984).

In practice this means that when we train a descriptive panel, we must be careful to facilitate meaningful concept formation by exposing the panel to as many standards as feasible. However, if the concept boundaries are very clear and narrow (for example, sweetness) a single standard may be adequate. Concept formation is improved when it occurs within the product category under study. For example, Sulmont et al. (1999) working with orange juice found that panels receiving spiked orange juice samples as reference standards were more discriminant and homogeneous than panelists receiving either a single reference standard for each attribute or three reference standards per attribute. In their case it seemed that multiple reference

standards actually had a negative effect on the panel performance. But, Murray et al. (2001) caution that reference standards external to the product category also have a role to play in concept formation. It is important to note that the use of reference standards does not necessarily eliminate contrast effects and the sensory scientist should keep that in mind.

If panelists are to use exact sensory descriptors' descriptions they must be trained. And untrained panelists frequently realize this when they are asked to evaluate products on attributes for which they have no clear concept. Armstrong et al. (1997) quote one of their untrained panelist: "I would rather we sat down and decided on what certain words and descriptions meant." The goal is for all panelists to use the same concepts and to be able to communicate precisely with one another; in other words the training process creates a "frame of reference" for the panel as a group (Murray et al., 2001). Thus, almost as an a priori assumption, descriptive analysis requires precise and specific concepts articulated in carefully chosen scientific language. The language used by consumers to describe sensory characteristics is almost always too imprecise and non-specific to allow the sensory specialist to measure and understand the underlying concepts in a way that will provide meaningful data.

Concept formation and definition can be illustrated as follows. In the United States and most Western countries our everyday concepts of colors are very similar, because we are taught as children to associate certain labels with certain stimuli. In other words, if a child says that the leaves of an oak tree are blue, the child will be told that the leaves are green. If the child persists in misnaming the color then the child would be tested for vision and/or other problems. Color is thus a well-structured concept for most individuals and possesses a widely understood scientific language for its description. However, with other sensory attributes such as flavor this is not true. In our culture we rarely describe the flavor of a food in precise terms. We usually say things like "the freshly baked bread smells good" or "the cough syrup tastes bad." There are charts with standard colors with coded labels (for example, the Munsell Book of Colors) but for taste, odor, and texture there is no "Munsell Book" and thus when we want to do research on these concepts we need to precisely define (preferably with reference standards) the scientific language used to describe the sensory sensations associated with the products studied.

When selecting terms (descriptors) to describe the sensory attributes of products we must keep the several desirable characteristics of descriptors in mind (Civille and Lawless, 1986). The desirable characteristics discussed by Civille and Lawless and others are listed in order of approximate importance in Table 10.1. We will consider each of these characteristics in turn. The selected descriptors should discriminate among the samples; therefore, they should indicate perceived differences among the samples. Thus, if we are evaluating cranberry juice samples and all the samples are the exact same shade of red then "red color intensity" would not be a useful descriptor. On the other hand, if the red color of the cranberry juice samples differs, due to processing conditions for example, then "red color intensity" would be an appropriate descriptor.

Table 10.1 Desirable characteristics that should be remembered when choosing terms for descriptive analysis studies (in order of importance)

Discriminate	More important
Non-redundant	
Relate to consumer acceptance/rejection	
Relate to instrumental or physical measurements	
Singular	
Precise and reliable	
Consensus on meaning	
Unambiguous	
Reference easy to obtain	
Communicate	
Relate to reality	Less important

The chosen term should be completely non-redundant with other terms; an example of redundancy is when the panelists are evaluating a steak and they are asked to rate both the perceived tenderness and the toughness of the meat (Raffensperger et al., 1956) since they both indicate the same concept in meat. It would be much better to decide that either the term "toughness" or the term "tenderness" should be used in the evaluation of the meat samples. Additionally the terms should be orthogonal. Orthogonal descriptors are not correlated with each other. Non-orthogonal descriptors overlap; for example, asking a panel to score the "red fruit intensity" of a Pinot noir wine and to score "cherry intensity" would be asking them to score non-orthogonal terms. It is very confusing, de-motivating, and mentally frustrating to the panelists when they are asked to score

redundant and non-orthogonal terms. Sometimes, it is impossible to entirely eliminate term redundancy and to ensure that all terms are orthogonal. For example, in a study describing differences among vanilla essences, Heymann (1994a) trained a panel to evaluate both butterscotch odor and sweet milk flavor. The panel was convinced that these two terms described different sensations. Yet, during the data analysis it became clear from the principal component analysis that the two terms were redundant and that they exhibited a great deal of overlap. But, it is possible that while these terms were correlated in this product category, they may not be for another set of products!

Panelists often have preconceived notions about which terms are correlated and which are not. During training it is often necessary to help panelists "decorrelate" terms (Civille and Lawless, 1986; Lawless and Corrigan, 1993). In texture analysis panelists frequently cannot grasp the differences between denseness and hardness, since these terms are correlated in many foods but not all. Some foods are dense but not hard (cream cheese, refrigerated butter) and other foods are hard but not dense (American malted milk bars, refrigerated "aerated chocolate bars" in the United Kingdom). Exposing panelists to these products would help de-correlate these terms, allowing the panel to understand that the two terms do not always have to vary together.

The data from descriptive analyses are often used to interpret consumer hedonic responses to the same samples. Therefore, it is very helpful if the descriptors used in the descriptive analysis can be related to concepts that lead consumers to accept or reject the product. In a sensory profiling of aged natural cheeses the panel trained by Heisserer and Chambers (1993) chose to use the term "butyric acid" (a chemical name) instead of the panel's consensus term for the sensory odor impression, namely "baby vomit." In this case the term that they discarded might have been more helpful in relating consumer acceptance or rejection of the cheese than the more precise chemical term chosen. Also, the ideal descriptors can be related to the underlying natural structure (if it is known) of the product. For example, many terms associated with the texture profile are tied to rheological principles (Szczesniak et al., 1963). It is also possible to use terms that are related to the chemical nature of the flavor compounds found in the product. For example, Heymann and Noble (1987) used the term "bell pepper" to describe the odor sensation in Cabernet sauvignon wines associated with the chemical 2-methoxy-3-isobutyl pyrazine. The pyrazine odorant is present in Cabernet sauvignon wines and it is also the impact compound for bell pepper aroma. The use of "butyric acid" by Heisserer and Chambers (1993) to describe a specific odor in aged cheese is tied to the compound probably responsible for the odor.

Descriptors should be singular rather than combinations of several terms. Combination or holistic terms such as creamy, soft, clean, fresh are very confusing to panelists. Integrated terms may be appropriate in advertising but not in sensory analysis. These terms should be broken down into their elemental, analytical, and primary parts. For example, a number of scientists have found that creaminess perception is a function of smoothness, viscosity, fatty mouth feel, and cream flavor (see Frøst and Janhøj, 2007, for an excellent overview). A study involving creaminess will likely be more easily interpreted and understood if most or all of these terms are examined. Also, the term acrid is a combination of aroma and tactile sensations (Hegenbart, 1994), and panelists should be trained to evaluate the components of acrid rather than the integrated term itself. The term soft, as used with fabrics, is a combination of compressibility, springiness, smoothness to touch, and a lack of crisp edges when folded. The problem with compound descriptors like creamy is that they are not actionable. Product developers do not know what to fix if the data indicate that there is a problem with this descriptor. Do they change the viscosity? The particle size? The aroma? It is possible that the term is not weighted similarly by all panelists; some may emphasize the thickness concept and others the cream aroma which often vary independently, thus "muddling up" the analysis. This is clearly not a good state of affairs for a descriptive analysis panel.

Suitable descriptors are ones that can be used with precision and reliability by the panelists. Panelists should fairly easily agree on the meaning of a specified term, the term should thus be unambiguous. They should be able to agree on the prototypical examples related to the descriptor and they should agree on the boundaries of the descriptor. Using reference standards to signify these boundaries is encouraged. It simplifies the life of the panel leader if the physical reference standards for the descriptor are easy to obtain. However, difficulties in obtaining physical reference standards should not prevent the panel leader

or the panelists from using terms that are ideal in every other way.

The chosen descriptors should have communication value and should not be jargon. In other words, the terms should be understandable to the users of the information obtained in the study and not only to the descriptive panel and their leader. It is also helpful if the term had been used traditionally with the product or if it can be related to the existing literature. The reference standards associated with each descriptor have a two-fold purpose: to align the underlying concepts for the panelists and to act as "translation" devices for users of the information obtained from the study. Giboreau et al. (2007) stressed that circularity should be avoided in defining sensory descriptors, for example, "noisy" should not be defined as "that which makes noise" but rather as "that which produces sound when it is bitten." These authors also stress that reference standards would increase the utility of these definitions and that definitions should be exact substitutes for the defined terms. An example would be "This piece of meat is very tough" and substituting the definition for "tough" one would say "This piece of meat is very difficult to chew."

Krasner (1995) working with water taints showed that some reference standards, for example, a hypochlorite solution for chlorine odor in water or a piece of boiled rubber hose for a rubbery smell, were distinctive and a large percentage of panelists agreed on the odor. Other chemicals were not successful as reference standards, for example, hexanal evoked a grassy odor descriptor from about 24% of his panelists and a lettuce aroma descriptor from 41% of the panelists with the rest divided between celery, olives, tobacco smoke, and old produce. We are of the opinion that this occurs relatively frequently with single chemical compounds.

The use of multiple reference standards for a single concept enhances learning and use of the concept (Ishii and O'Mahony, 1991). Additionally, panel leaders with a broad sensory reference base facilitate learning. For example, panelist responses to the odor of oil of bitter almonds may include descriptors such as almond, cherry, cough drops, Amaretto, and Danish pastries. All of these descriptors refer to the underlying benzaldehyde character in all these products. In another study the panelists may state that the product reminds them of cardboard, paint, and linseed oil. The experienced panel leader will realize that all these terms are descriptive of the sensation associated with the oxidation of lipids and fatty acids. It is also helpful if the panel leader has background knowledge of the product category.

10.4 Descriptive Analysis Techniques

In the following section we will review the major approaches and philosophies of descriptive analysis techniques. Reviews can be found in Amerine et al. (1965), Powers (1988), Einstein (1991), Heymann et al. (1993), Murray et al. (2001), Stone and Sidel (2004), and Meilgaard et al. (2006). Additionally, Muñoz and Civille (1998) clearly explained some of the philosophical differences with respect to panel training and scale usage among the different techniques.

10.4.1 Flavor Profile®

In its original incarnation the Flavor Profile (FP) is a qualitative descriptive test. The name and the technique were trademarked to Arthur D. Little and Co., Cambridge, MA. This technique was developed in the late 1940s and early 1950s at Arthur D. Little by Lören Sjostrom, Stanley Cairncross, and Jean Caul. FP was first used to describe complex flavor systems measuring the effect of monosodium glutamate on flavor perception. Over the years FP was continually refined. The latest version of FP is known as Profile Attribute Analysis (Cairncross and Sjöstrom, 1950; Caul, 1957, 1967; Hall, 1958; Meilgaard et al., 2006; Moskowitz, 1988; Murray et al., 2001; Powers, 1988; Sjöström, 1954).

Flavor profiling is a consensus technique. The vocabulary used to describe the product and the product evaluation itself is achieved by reaching agreement among the panel members. The FP considers the overall flavor and the individual detectable flavor components of a food system. The profile describes the overall flavor and the flavor notes and estimates the intensity of these descriptors and the amplitude (overall impression). The technique provides a tabulation of the perceived flavors, their intensities, their order of perception, their aftertastes, and their overall impression (amplitude). If the panelists are trained appropriately this tabulation is reproducible.

Using standardized techniques of preparation, presentation, and evaluation, the four to six judges are trained to precisely define the flavors of the product category during a 2- to 3-week program. The food samples are tasted and all perceived notes are recorded for aroma, flavor, mouth feel, and aftertaste. The panel is exposed to a wide range of products within the food category. After this exposure the panelists review and refine the descriptors used. Reference standards and definitions for each descriptor are also created during the training phase. Use of appropriate reference standards improves the precision of the consensus description. At the completion of the training phase the panelists have defined a frame of reference for expressing the intensities of the descriptors used.

The samples are served to the panelists in the same form that they would be served to the consumer. Thus, if the panel was studying cherry pie fillings the filling would be served to the panel in a pie.

Originally, the intensities of the perceived flavor notes were rated on the following scale (this scale has subsequently been expanded with up to 17 points including the use of arrows, ½'s, or plus and minus symbols):

Rating	Explanation
0	Not present
)(	Threshold or just recognizable
1	Slight
2	Moderate
3	Strong

The order in which the flavor notes are perceived is also indicated on the tabulated profile. The aftertaste is defined as one or two flavor impressions that are left on the palate after swallowing. The panel rates the aftertaste intensities 1 min after the product is swallowed.

The amplitude is the degree of balance and blending of the flavor. It is not supposed to be indicative of the overall quality of the product nor is it supposed to include the panelists' hedonic responses to the product. Proponents of FP admit that it is very difficult for novice panelists to divorce their hedonic responses from the concept of amplitude. However, panelists do reach an understanding of the term with training and exposure to the FP method and the product category. The amplitude is defined as an overall impression of balance and blending of the product. In a sense, the amplitude is not to be understood, just to be experienced. For example, heavy cream, when whipped, has a low amplitude; heavy cream whipped with the addition of some sugar has a higher amplitude; and heavy cream whipped with the addition of some sugar and vanilla essence has a much higher amplitude. Usually, FP panelists determine the amplitude before they concentrate on the individual flavor notes of the product. However, the amplitude may be placed last in the tabular profile. The following scale is used to rate amplitude:

Rating	Explanation
)(	Very low
1	Low
2	Medium
3	High

The panel leader derives a consensus profile from the responses of the panel. In a true FP this is not a process of averaging scores, but rather that the consensus is obtained by discussion and re-evaluation of the products by the panelists and panel leader. The final product description is indicated by a series of symbols. As described earlier, these are a combination of numerals and other symbols that are combined by the panelists into potentially meaningful patterns, whether as a descriptive table (Table 10.2) or as a graphic, the "sunburst."

The "sunburst," which is not used currently, was a graphical representation of FP results (Cairncross and Sjöstrom, 1950). A semi-circle indicates the threshold intensity and the radiating line lengths indicated the consensus intensity of each attribute evaluated. The order in which various characteristics "emerge" from the sample is noted by the order (from left to right) on the graph. While these symbols can be used to describe the product, it is impossible to analyze the data obtained in this way by conventional statistical procedures. Therefore, the FP is classified as a qualitative descriptive technique.

With the introduction of numerical scales, between 1 and 7 points, (Moskowitz, 1988), the Flavor Profile was renamed the Profile Attribute Analysis (PAA). Data derived from PAA may be statistically analyzed but it is also possible to derive a FP-type consensus description. The use of numerical scales allows researchers employing this method to use statistical techniques to facilitate data interpretation. PAA is more quantitative than FP (Hanson et al., 1983).

Table 10.2 Example of the consensus result of a flavor profile study. Composite flavor profile for turkey patties with 0.4% added phosphate

Flavor	Attributes
	Intensity[a]
Protein	2–
Meat identity	1
Serumy	1
[pause]	
Metallic (aromatic and feel)	1+
(Carries through)	1–
Poultry	1+
Brothy	1–
[lag]	
Turkey	1
Organ meat	1–
Metallic (aromatic and feel)	1
Bitter	)(
Aftertaste	Intensity[a]
Metallic feel	2–
Poultry	1–
Other[b]	
Turkey	)(+
Organ meat	)(+

Adapted from Chambers et al. (1992)
[a]Scale:)(= threshold, 1 = slight, 2 = moderate, 3 = strong
[b]"Other" characteristics in the aftertaste were not found by the entire panel

Syarief and coworkers (1985) compared flavor profile results derived through consensus with flavor profile results derived by calculating mean scores. The mean score results had a smaller coefficient of variation than the consensus results and the principal component analysis (PCA) of the mean score data accounted for a higher proportion of the variance than the PCA of the consensus scores. Based on these criteria the authors concluded that the use of mean scores gave superior results to that of consensus scores. Despite these results, some practitioners still use both the FP and PAA as a consensus technique.

Proponents of FP state that the data are accurate and reproducible if the panelists are well trained. The necessity for vocabulary standardization among panelists cannot be overestimated. Detractors of these procedures complain that the derived consensus may actually be the opinion of the most dominant personality in the group or the panel member perceived to have the greatest authority, often the panel leader. Advocates of the techniques counter that with proper training the panel leader will prevent this from occurring. Additionally, champions of the method maintain that a trained FP panel produces results rapidly. Proper training is critical when using these techniques successfully.

True FP resists most attempts for mathematical characterization of the data. Usually a series of symbols must be interpreted using intuition and experience on the part of the researcher. PAA, on the other hand, can be analyzed using parametric techniques such as analysis of variance and suitable means separation procedures. Currently, the FP technique is used extensively in the evaluation of water, probably because water utilities usually only have three to four people to troubleshoot taste and odor complaints (AWWA, 1993; Bartels et al., 1986, 1987; Ömür-Özbek and Dietrich, 2008).

10.4.1.1 Flavor Profile Judge Selection

Flavor Profile judges should be screened for long-term availability. It takes time, effort, and money to train a panel and the panelists should make a commitment to be available for years, if possible. It is not unusual to find FP panelists who have served on the same panel for more than 10 years. Potential panelists should have a keen interest in the product category and it is helpful if they have some background knowledge on the product type. These panelists should be screened to have normal odor and taste perceptions. Panelists are screened for normal acuity using solutions and pure diluted odorants (see Chapter 2). They should be very articulate and sincere with an appropriate personality (not timid or overly aggressive).

The panel leader is an active participant in both the language development and evaluation phases of the study. The panel leader must moderate the interactions between panelists, leading the entire group toward some unanimity of opinion. It is clear that the key element in a FP panel is the panel leader. This person coordinates the sample production, directs the panel evaluations, and finally verbalizes the consensus conclusions of the entire panel. The panel leader will often resubmit samples until reproducible results are obtained. Therefore, the panel leader should be especially articulate and knowledgeable about the product type. This person will also be responsible for communication with the panel and preparation of samples and reference standards. The panel leader should also

be infinitely patient, socially sensitive, and diplomatic since he/she will be responsible for moving the panel to a consensus description of the product.

10.4.2 Quantitative Descriptive Analysis®

Quantitative Descriptive Analysis (QDA) was developed during the 1970s to correct some of the perceived problems associated with the Flavor Profile analysis (Stone and Sidel, 2004; Stone et al., 1974). In contrast to FP and PAA, the data are not generated through consensus discussions, panel leaders are not active participants, and unstructured line scales are used to describe the intensity of rated attributes. Stone et al. (1974) chose the linear graphic scale, a line that extends beyond the fixed verbal end points, because they found that this scale may reduce the tendency of panelists to use only the central part of the scale avoiding very high or very low scores. Their decision was based in part on Anderson's studies (1970) of functional measurement in psychological judgments. As with FP, QDA has many advocates and the technique has been extensively reviewed (Einstein, 1991; Heymann et al., 1993; Meilgaard et al., 2006; Murray et al., 2001; Powers, 1988; Stone and Sidel, 2004; Stone et al., 1980; Zook and Wessman, 1977).

During QDA training sessions, 10–12 judges are exposed to many possible variations of the product to facilitate accurate concept formation. The choice of range of samples is dictated by the purpose of the study and, similar to FP, panelists generate a set of terms that describe differences among the products. Then through consensus, panelists develop a standardized vocabulary to describe the sensory differences among the samples. The panelists also decide on the reference standards and/or verbal definitions that should be used to anchor the descriptive terms. Actual reference standards are only used about 10% of the time; usually, only verbal definitions are used (Murray et al., 2001). In addition, during the training period the panel decides the sequence for evaluating each attribute. Late in the training sequence, a series of trial evaluations are performed. This allows the panel leader to evaluate individual judges based on statistical analysis of their performance relative to that of the entire panel. Evaluations of panelist performance may also be performed during the evaluation phase of the study.

Panelists begin training by generating a consensus vocabulary. During these early meetings, the panel leader acts *only* as a facilitator by directing discussion and supplying materials such as reference standards and product samples as required by the panel. The panel leader does not participate in the final product evaluations.

Unlike FP, QDA samples may not be served exactly as seen by the consumer. For example, if a Flavor Profile panel is to evaluate pie crusts, they would receive samples of pie crust filled with a standard pie filling. The QDA philosophy states that the pie filling could affect the discrimination of the crust samples. However, a case could also be made that crust baked without filling may perform differently than crust baked with filling. Depending on the situation, the QDA panelists may receive two different pie crust samples, one baked without filling and the other baked with filling, which was removed before the panelists received the crust samples.

The actual product evaluations are performed by each judge individually, usually while seated in isolated booths. Standard sensory practices such as sample coding, booth lighting, expectorating, and rinsing between samples are used for the evaluation phase. A 6 in. graphic line scale anchored with words generated by the panel is used (Fig. 10.1).

The resulting data can be analyzed statistically using analysis of variance and multivariate statistical techniques. It is necessary for judges to replicate their judgments, up to six times in some cases, to allow the sensory scientist to check the consistency of the individual panelists and of the whole panel.

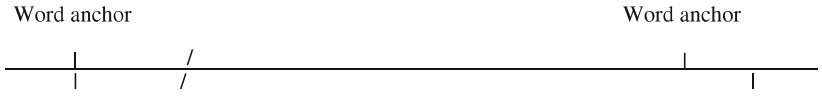

Fig. 10.1 An example of the QDA graphic line scale. The mark made by the panelist is converted to a numerical value by measuring from the *left end of the line*.

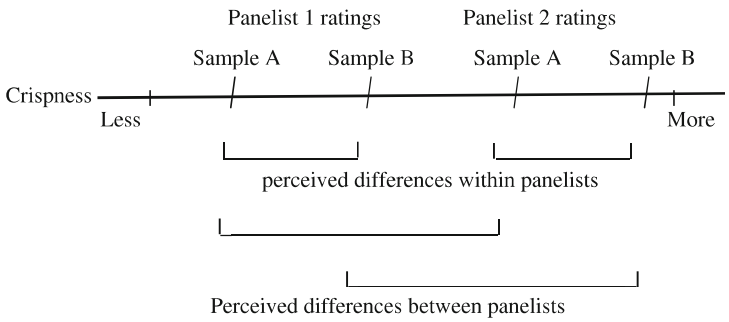

Fig 10.2 Different uses of the line scale by panelists that are calibrated *relative* to each other. All ratings were plotted on the same line scale for illustration purposes.

Replications also allow one-way analysis of variance of individual panelists across products. This allows the sensory specialist to determine whether the panelists can discriminate or need more training. The number of repeat judgments is somewhat product dependent and should be decided before the study is initiated. Studies where repeated evaluations are not performed should be viewed with extreme caution.

QDA may be used to completely describe the sensory sensations associated with a product from initial visual assessment to aftertaste, or panelists may be instructed to focus on a narrow range of attributes such as texture descriptors. However, limiting the range of attributes evaluated may lead to the "dumping" effect (see Chapter 9). This effect is especially important when a conspicuous sensory attribute that varies across the samples was omitted from the ballot. When this occurs panelists will, probably sub-consciously, express their frustration by modulating the scores on some of the scales used in the study. For this reason, the sensory scientist should be extremely careful about restricting the type and number of descriptors used in a descriptive analysis study. Sometimes, simply adding a scale labeled "other" can prevent this effect and if the panelists are allowed to describe the "other" characteristic valuable information may also be obtained.

During training, one of the challenges faced by the panel leader is how to help judges sort out the individual intensity characteristics of a product from overall impressions of quality or liking (Civille and Lawless, 1986). All descriptive evaluations should only be based on perceived intensities and should be free of hedonic responses.

Despite the extensive training employed in this method, most researchers assume that judges will use different parts of the scale to make their determinations. Thus, the absolute scale values are not important. It is the relative differences among products that provide valuable information. For example, Judge A scores the crispness of potato chip sample 1 as an 8, but Judge B scores the same sample as a 5; this does not mean that the two judges are not measuring the same attribute in the same way, but may mean that they are using different parts of the scale (Fig. 10.2). The relative responses of these two judges on a second different sample (say 6 and 3, respectively) would indicate that the two judges are calibrated with respect to the *relative* differences between the samples. Judicious choices of statistical procedures such as dependent *t*-tests and ANOVA allow the researcher to remove the effect of using different parts of the scale.

QDA training often takes less time than that required by FP. Consensus by personality domination, a potential problem with FP, is unlikely to occur since individual judgments are used in the data analysis. In addition, QDA data are readily analyzed by both univariate and multivariate statistical techniques. Statistical procedures such as multivariate analysis of variance, principal component analysis, factor analysis, cluster analysis have found application in the analysis of data generated by QDA-type procedures (Martens and Martens, 2001; Meullenet et al., 2007; Piggott, 1986). Graphical presentations of the data often involve the use of "cobweb" graphs (polar coordinate graphs a.k.a. radar plots, Fig. 10.3).

There is some argument about the assumption of normal distribution of the data set and hence the use of parametric statistics such as analysis of variance, *t*-tests. A few authors feel that non-parametric statistical treatment of the data is required (O'Mahony, 1986; Randall, 1989), but this appears to be a minority opinion.

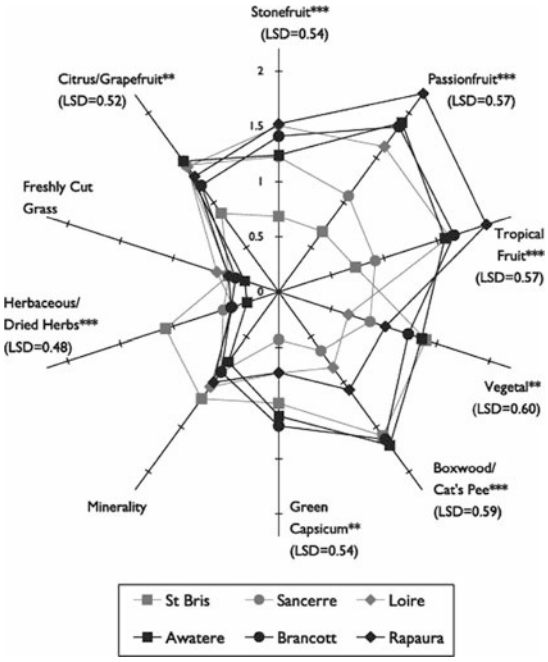

Fig. 10.3 An example of a cobweb or radar plot of Descriptive Analysis data. The data are from aroma profile descriptive analysis of Sauvignon blanc wines as a function of country of origin (France or New Zealand) and sub-region (France: Saint Bris, Sancerre, Loire; New Zealand: Awatere, Brancott, Rapaura). For each sensory attribute the perceived mean intensity increases outward from the center point. Sub-region means differing by more than the LSD value for that attribute differ in a Fisher's LSD multiple comparison test (Parr et al., 2009, used with permission).

The ease of data analysis using QDA may actually be considered one of the problems of the technique. The tendency to use the scales as absolute measures of an attribute rather than as a tool to see relative differences between samples is very common. Returning to the potato chip example, a decision may be made that samples scoring less than 5 on the crispness scale are not acceptable for sale. As we saw, Judge B's crispness intensity of 5 was very different from Judge A's 5. By extension, we can see that if the entire panel used the upper end of the scale, no samples would be considered unacceptable by this criterion. If another panel, analyzing the same samples, uses only the lower end of the scale, no sample is acceptable. The QDA data must therefore be viewed as *relative* values and not as absolutes. QDA studies should therefore be designed to include more than one sample and/or a benchmark or standard product as often as possible.

QDA has been extensively used, but often the experiments are not designed exactly as described by Stone and Sidel (2004). The relative simplicity of the technique allows it to be adapted in many different ways. However, any adaptation invalidates the use of the name QDA to describe the procedure.

Advantages cited by advocates of QDA include the ideas that the panelists perform independent judgments and that results are not consensus derived. Additionally, the data are easily analyzed statistically and graphically represented. Panel language development is free of panel leader influences and is, in general, based on consumer language descriptions. QDA suffers from the same disadvantage as FP, since in both cases the panels must be trained for the specific product category. Many US food companies maintain separate panels for their many product categories. This is very expensive and may limit the use of this technique by smaller firms. Unlike FP, the QDA results

do not necessarily indicate the order of perception of sensations. However, the panel could be instructed to list the descriptors on the ballot in sequence of appearance, if that is needed to meet the objective of the study. Additionally, as indicated above, the results are relative and not absolute, since panelists may use different scale ranges.

10.4.2.1 Selection of Judges for Quantitative Descriptive Analysis

Similar to FP judges, Quantitative Descriptive Analysis (QDA) panelists should be screened for long-term availability. As with FP, it takes time, effort, and money to train a panel and the panelists should make a commitment to be available for years, if possible. This becomes a management support issue when the panelists are selected from within company since these employees may spend substantial time away from their main jobs. These panelists are screened for normal odor and taste perceptions using actual products from the category. The potential panelists should be very articulate and sincere.

Unlike FP, the panel leader is not an active participant in either the language development or the evaluation phases of the study. The panel leader acts as a facilitator only and does not lead or direct the panel. This person will be responsible for communication with the panel and preparation of samples and reference standards.

10.4.3 Texture Profile®

The Texture Profile was created by scientists working for General Foods during the 1960s and was subsequently modified by several sensory specialists (Brandt et al., 1963; Civille and Liska, 1975; Muñoz, 1986; Szczesniak, 1963, 1966, 1975; Szczesniak et al., 1963). The goal of the Texture Profile (TP) was to devise a sensory technique that would allow the assessment of all the texture characteristics of a product, from first bite through complete mastication, using engineering principles. The creators based the Texture Profile on the concepts pioneered by developers of the Flavor Profile. The texture profile was defined by Civille and Liska (1975) as

> the sensory analysis of the texture complex of a food in terms of its mechanical, geometrical, fat and moisture characteristics, the degree of each present and the order in which they appear from fist bite through complete mastication (p. 19).

The Texture Profile uses a standardized terminology to describe the textural characteristics of any product. Specific characteristics are described by both their physical and sensory aspects. Product-specific terms to be employed are chosen from the standardized terminology to describe the texture of a specific product. Definitions and order of appearance of the terms are decided through consensus by the TP panelists. Rating scales associated with the textural terms are standardized (Table 10.3).

Table 10.3 Example texture profile hardness[a] scale

Scale value	Product	Sample size	Temperature	Composition
1.0	Cream cheese	½" cube	40–45°C	Philadelphia cream cheese (Kraft)
2.5	Egg white	¼" cube	Room	Hard-cooked, 5 min
4.5	American cheese ½"' cube	40–45°C	Yellow pasteurized cheese (Land O Lakes)	
6.0	Olive	1 piece	Room	Stuffed, Spanish olives with pimentos removed (Goya Foods)
7.0	Frankfurter[b]	½" slice	Room	Beef Franks, cooked for 5 min in boiling water (Hebrew National Kosher Foods)
9.5	Peanut	1 piece	Room	Cocktail peanuts in vacuum tin (Planters, Nabisco Brands)
11.0	Almond	1 piece	Room	Shelled almonds (Planter, Nabisco Brands)
14.5	Hard candy	1 piece	Room	Life Savers (Nabisco Brands)

Adapted from Muñoz (1986)

[a]Hardness is defined as the force required to bite completely through sample placed between molar teeth

[b]Area compressed with molars is parallel to cut

Within each scale, the full range of a specific parameter is anchored by products having the specific characteristic as a major component. The reference product must be instrumentally evaluated to determine whether they conform to the intensity increments for the specified scale. The reference scales anchor both the range and the concept for each term (Szczesniak et al., 1963). For example, the hardness scale (Table 10.3) measures the compression force applied to the product between the molars. Note that the different foods (cream cheese, cooked egg white, cheese, olives, wieners, peanuts, raw carrots, almonds, and hard candies) used as reference points in the TP hardness scale increase in intensity from cream cheese to candy. However, these products alternately may shear, shatter, or compress when the compression force is applied. Thus when using the hardness reference scale panelists must understand that although all these products vary in a specific and definable dimension, hardness, they do not necessarily react in the same way to an applied compressive force.

It is crucial to the success of the TP that the frame of reference for all panelists be the same. All panelists must receive the same training in the principles of texture and TP procedures. Preparation, serving, and evaluation of samples should be rigidly controlled. Panelists should also be trained to bite, chew, and swallow in a standardized way. Usually, during panel training the panelists are first exposed to the Szczesniak (1963) classification of textural characteristics. They are subsequently exposed to a wide variety of food products and reference scales. During the third phase, the panelists refine their skills in recognizing, identifying, and quantifying degrees within each textural characteristic in a specific food category. This normally takes several weeks of daily training sessions but this may be worthwhile. Otremba et al. (2000), working with beef muscles, found that the extensive training led to greater consistency and accuracy.

The Texture Profile has been applied to many specific product categories including breakfast cereals, rice, whipped toppings, cookies, meat, snack foods. However, in many cases the experimenters will state that they had used TP in their studies but careful analysis of their methodology reveals that the exacting requirements of true TP were not adhered to during these studies. Often panelists are not trained using the standardized methodology to the degree recommended by the original proponents of this technique.

10.4.4 Sensory Spectrum®

Gail Civille, while working at General Foods in the 1970s, became an expert using the Texture Profile. She, subsequently, created the Sensory Spectrum technique using many of the ideas inherent to the Texture Profile. The Sensory Spectrum procedure is a further expansion of descriptive analysis techniques. The unique characteristic of the Spectrum approach is that panelists do not generate a panel-specific vocabulary to describe sensory attributes of products, but that they use a standardized lexicon of terms (Civille and Lyon, 1996). The language used to describe a particular product is chosen a priori and remains the same for all products within a category over time. Additionally, the scales are standardized and anchored with multiple reference points. The panelists are trained to use the scales identically; because of this, proponents of the Spectrum method state that the resultant data values are absolute. This means that it should be possible to design experiments that include only one sample and to compare the data from that sample with data derived in a different study. This philosophy suggests that since each panel is a unique group, allowing panels to generate their own consensus terms may lead to misleading results when attempting to apply the findings to a generalized population. The proponents of the method state that the descriptors used for the Spectrum method are more technical than the QDA descriptors. According to Sensory Spectrum users, QDA terms are generated by the panelists themselves and they are more likely to be related to consumer language. Reviews of the Spectrum method have been provided by Powers (1988), Murray et al. (2001), and Meilgaard and coworkers (2006).

Panelist training for the Spectrum method is much more extensive than QDA training and the panel leader has a more directive role than in QDA. As in QDA, the judges are exposed to a wide variety of the products in the specific product category. As in the Texture Profile, the panel leader provides extensive information on the product ingredients. The underlying chemical, rheological, and visual principles are explored by the panelists and the relationships between these principles and sensory perceptions of the products are considered. Similar to the Texture Profile the panelists are provided word lists (called lexicons by Sensory Spectrum) that may be used to describe perceived

sensations associated with the samples. The ultimate goal is to develop an "…'expert panel' in a given field, … [to] demonstrate that it can use a concrete list of descriptors based on an understanding of the underlying technical differences among the attributes of the product" (Meilgaard et al., 2006). Additionally, panelists are supplied with reference standards. For attributes, specific singular references as well as standards in combination with a few other attributes are provided. An example would be vanilla and vanilla in milk and/or cream (Muñoz and Civille, 1998).

Panelists use intensity scales that are numerical, usually 15-point scales, and *absolute* also known as *universal* (Table 10.4; Muñoz and Civille, 1997). Civille (April 1996, personal communication) states that the scales are created to have equi-intensity across scales. In other words, a "5" on the sweetness scale is equal in intensity to a "5" on the salty scale and this is even equal in intensity to a "5" on the fruity scale (Table 10.5). Civille (April 1996, personal communication) says this goal has been achieved for fragrance, aroma, and flavor scales but not for texture scales. We are somewhat skeptical of this equi-intensity claim since there are no published data to support it. However, the concept of cross-modal matching may make the above claim reasonable for light and tones, tastants (sweetness and sourness), but it may not be reasonable for sweetness and hardness or fruitiness and chewiness (Stevens, 1969; Stevens and Marks, 1980; Ward, 1986).

Also, the stability of the absolute scale is not clear. Olabi and Lawless (2008) found contextual shifting in the 15-point scale even after extensive training.

As with the Texture Profile, scales are anchored by a series of reference points. In this schema at least two and preferably three to five references are recommended. The reference points are chosen to represent different intensities on the scale continuum. The reference points are used to precisely calibrate the panelists in the same way as pH buffers calibrate a pH meter. The panelists are "tuned" to act like true instruments.

Table 10.4 Example of aromatic reference samples used for spectrum scales

Descriptor	Scale value[a]	Product
Astringency	6.5	Tea bags soaked for 1 h
	6.5	Grape juice (Welch's)
Caramelized sugar	3.0	Brown Edge Cookies (Nabisco)
	4.0	Sugar Cookies (Kroger)
	4.0	Social Tea Cookies (Nabisco)
	7.0	Bordeaux Cookies (Pepperidge Farm)
Egg	5.0	Mayonnaise (Hellmann's)
Egg flavor	13.5	Hard boiled egg
Orange complex	3.0	Orange Drink (Hi-C)
	6.5	Reconstituted frozen orange concentrate (Minute Maid)
	7.5	Freshly squeezed orange juice
	9.5	Orange concentrate (Tang)
Roastedness	7.0	Coffee (Maxwell House)
	14.0	Espresso coffee (Medaglia D'Oro)
Vanilla	7.0	Sugar Cookies (Kroger)

Adapted from Meilgaard et al. (2006)
[a]All of the above scales run from 0 to 15

Table 10.5 Intensity values used for spectrum scales assigned to the four basic tastes in assorted products

Descriptor	Scale value[a]	Product
Sweet	2.0	2% sucrose-water solution
	4.0	Ritz cracker (Nabisco)
	7.0	Lemonade (Country Time)
	9.0	Coca Cola Classic
	12.5	Bordeaux Cookies (Pepperidge Farm)
	15.0	16% sucrose-water solution
Sour	2.0	0.05% citric acid-water solution
	4.0	Natural apple sauce (Motts)
	5.0	Reconstituted frozen orange juice (Minute Maid)
	8.0	Sweet pickle (Vlasic)
	10.0	Kosher dill pickle (Vlasic)
	15.0	0.20% citric acid-water solution
Salt	2.0	0.2% sodium chloride-water solution
	5.0	Salted soda cracker (Premium)
	7.0	American cheese (Kraft)
	8.0	Mayonnaise (Hellman's)
	9.5	Salted potato chips (Frito-Lay)
	15.0	1.5% sodium chloride-water solution
Bitter	2.0	Bottled grapefruit juice (Kraft)
	4.0	Chocolate bar (Hershey)
	5.0	0.08% caffeine-water solution
	7.0	Raw endive
	9.0	Celery seed
	10.0	0.15% caffeine-water solution
	15.0	0.20% caffeine-water solution

Adapted from Meilgaard et al. (2006)
[a]All the above scales run from 0 to 15

After training, all panelists must use the scales in an identical fashion. Thus, they should all score a specific attribute of a specific sample at the same intensity. Testing is performed in isolated booths, using typical sensory practices.

The principal advantage claimed for the Spectrum method should be apparent after reading the discussion of the QDA procedure. In QDA, judges frequently use the scales provided in idiosyncratic but consistent ways. In contrast to the QDA, the Spectrum method trains all panelists to use the descriptor scales in the same way. Thus, scores should have absolute meaning. This means that mean scores could be used to determine if a sample with a specified attribute intensity fits the criterion for acceptability irrespective of panel location, history, or other variables. This has obvious advantages to organizations wishing to use a descriptive technique in routine quality assurance operations or in multiple locations and facilities.

Disadvantages of the procedure are associated with the difficulties of panel development and maintenance. Training of a Spectrum panel is usually very time consuming. Panelists have to be exposed to the samples and understand the vocabulary chosen to describe the product. They are asked to grasp the underlying technical details of the product and they are expected to have a basic understanding of the physiology and psychology of sensory perception. After all that, they must also be extensively "tuned" to one another to ensure that all panelists are using the scales in the same way. We are not sure that this level of calibration can be achieved in reality. In practice, individual differences among panelists related to physiological differences like specific anosmias, differential sensitivities to ingredients can lead to incomplete agreement among panelists. Theoretically, if panelists were in complete agreement one would expect the standard deviation (see Appendix) for any specific product–attribute combination to be close to zero. However, most Spectrum studies have attributes with non-zero standard deviations indicating that the panel is not absolutely calibrated. Civille (April 1996, personal communication) has stated that absolute calibration is feasible for most attributes but probably not for bitterness, pungency, and certain odor perceptions.

Data from the Spectrum technique are analyzed in a similar fashion to the QDA data. The deviation of mean values for particular attributes is of definite interest to the analyst, since these values can be directly related to the "tuning" or precision of the panel.

10.5 Generic Descriptive Analysis

QDA and Sensory Spectrum techniques have been adapted in many different ways. It is important to note, however, that any adaptations invalidate the use of the trademarked names "QDA" and "Sensory Spectrum." Unfortunately, it is often difficult to evaluate the effect that the myriad deviations from the standard methodologies have on the validity of the data. Academic researchers frequently employ the general guidelines of these methodologies to evaluate products. Table 10.6 shows the steps in conducting a generic descriptive analysis; these steps will be described in detail in the next sections. Additionally, some very interesting variations on the conventional generic descriptive analysis have been created and these will be discussed in Section 10.4.7.

10.5.1 How to Do Descriptive Analysis in Three Easy Steps

It is possible for any competent sensory scientist to perform a descriptive analysis study in three easy steps. These steps are train the judges, determine the judge reproducibility/consistency, and have the judges evaluate the samples. We will discuss each of these steps in more detail.

10.5.1.1 Training the Panelists

As we have seen with the QDA and Sensory Spectrum methods, there are two methods of judge training. The first is to provide the panelists with a wide range of products in the specific category. Panelists are asked to generate the descriptors and reference standards needed to describe differences among the products, usually by coming to some consensus. For simplicity we will call this "consensus training." The second method is to provide the panelists with a wide range of products within the category as well as a word

Table 10.6 Steps in conducting a generic descriptive analysis

1. Determine project objective: Is descriptive analysis the right method?
2. Establish products to be used with clients/researchers
3. Determine whether consensus or ballot training is most appropriate
4. Establish experimental design and statistical analyses
   a. Main effects and interactions for analyses of variance
   b. Multivariate techniques?
5. Select (and optionally, screen) panelists
   If choosing to do consensus training go to 6. If choosing to do ballot training go to 7
6. Consensus training
   a. During initial training sessions provide panelists with a wide range of products in the specific category
   b. Panelists generate descriptors (and ideas for reference standards)
   c. During subsequent training sessions panel leader provides potential reference standards as well as products
   d. Panelists reach consensus in terms of attributes, reference standards, and score sheet sequencing
7. Ballot training
   a. During initial training sessions provide panelists with a wide range of products in the specific category.
   b. Provide panelists with a word list (sample score sheet) and reference standards
   c. During subsequent training sessions panel leader provides reference standards as well as products
   d. Panelists indicate which attributes and reference standards from the word list should be used in the specific study. Panelists may also indicate sequence of attributes on score sheet
8. Once the training phase has been completed, panelists performance is checked
   a. A subset of samples are provided in duplicate (triplicate) under actual testing conditions
   b. Data are analyzed and any issues with reproducibility and/or attribute usage lead to additional training; testing may occur again after re-training.
9. Conduct study
10. Analyze and report data

list of possible descriptors and references that could be used to describe the products. We will refer to this method as "ballot training." In practice, both the consensus and the ballot methods have an application. However, keep in mind that Sulmont et al. (1999) found that panelists tended to perform better when trained by the "consensus" (trained by doing) rather than "ballot" (trained by being told) method. Frequently, however, a combination method is used. In the combination method panelists derive some descriptors on their own through consensus and others are added through suggestions by the panel leader or from word lists. The panel leader may also reduce redundant terms. In our laboratories the consensus method is usually used in research studies with the exception of meat research studies. For meat we tend to use the ballot method, mostly because a multitude of studies in the area has convinced us only a limited number of descriptors are readily applicable to meat. In contract work for US food and consumer products companies, we tend to use the combination method, since the client companies often have certain terms that they deem important. These will then be suggested by the panel leader, if the panelists do not use them spontaneously.

A typical sequence of "consensus training" sessions would be the following:

Initially, the panelists are exposed to the entire range of the products. They are asked to evaluate the sensory differences among the samples and to write down the descriptors that describe these differences. This occurs in silence. When all panelists complete this portion of the assignment, the panel leader asks each panelist to list the words used to describe each sample. During this phase of the training it is extremely important that the panel leader must be cautious not to lead or to judge any descriptor from any panelist. However, the panel leader may ask for clarification, if needed. Usually, the panelists themselves will begin to move toward initial consensus when they see the total list of descriptors elicited.

Subsequently, the panel leader should attempt to provide potential reference standards based on the initial consensus. These reference standards are chemicals, spices, ingredients, or products that can be used to help the panelists identify and remember the sensory attribute found in the samples evaluated (Rainey, 1986). In general, the panel leader should strive to use actual physical objects as the reference standards but in some cases precise written description may be used instead (Table 10.7). At the next session, the panelists are exposed to the samples again and asked to decide on the possible reference standards. If reference standards are not feasible, the panelists can also be asked to verbally define the specific descriptor. This refinement of the consensus list of descriptors, reference standards, and definitions continues until the panelists are satisfied that they have the best possible list and that everyone understands each term completely. Murray and Delahunty (2000) had their panelists determine

Table 10.7 Composition of reference standards for aroma and flavor evaluations. These reference standards were used in a descriptive study of vanilla essences from different geographic locations (Woods, 1995)

Aroma attribute	Composition
Smoky	20"' of binding twine lit with lighter, allowed to burn and then blown out, smell smoke
Scotch[a]	15 ml of 5% solution of J&B Justerini & Brooks Ltd., rare blended scotch whiskies (London, England)
Bourbon	15 ml of 5% solution of Walker's deluxe straight Bourbon Whiskey (Hiram Walker & Sons Co., Bardstown, KY)
Rum	15 ml of 5% solution of Bacardi Superior Puerto Rican Rum (Bacardi Corp., San Juan, Puerto Rico)
Almond	15 ml of 1.25% solution of McCormick® Pure Almond extract (McCormick & Co., Inc., Hunt Valley, MD)
Kahlua	15 ml of 1.25% solution of original Mexican Kahlua (Kahlua S.A., Rio San Joaquin, Mexico)
Medicinal	15 ml of 20% solution of Cepacol® mouthwash (Merrell Dow Pharmaceuticals, Inc., Cincinnati, OH)
Buttery	One piece of Lifesavers® Butter Rum candy (©Nabisco Foods, Inc., Winston-Salem, NC)
Creme Soda	15 ml of 2% solution of Shasta® creme soda (Shasta Beverages Inc., Hayward, CA)
Fruity	15 ml of 30% (5:1) solution of Welch's Orchard® apple-grape-cherry fruit juice cocktail frozen concentrate and Welch's® 100% white grape juice from concentrate (no sugar added) (©Welch's, Concord, MA)
Prune	One Sunsweet® medium prune (Sunsweet Growers, Stockton, CA)
Tobacco	One pinch of large size Beech-nut Softer & Moister chewing tobacco (©National Tobacco, Louisville, KY)
Earthy	19 g of black dirt from Missouri
Musty	**Verbally defined as "a damp basement"**
Nutty	2–3 Kroner® salted pistachios (shelled and cut into pieces) (Kroner Co., Cincinnati, OH)
Flavor attribute[b]	**Composition**
Amaretto	15 ml of 5% solution of Disaronno-Amaretto Originale (Illva Saronno, Saronno, Italy)
Sweet	Panelists were not provided with a reference, but were given a 2 and 6% solution of sugar water during training to anchor the scale
Fruity	15 ml of 5% (5:1) solution of Welch's Orchard® apple-grape-cherry fruit juice frozen concentrate and Welch's® 100% white grape juice from concentrate (no sugar added) (©Welch's, Concord, MA)
Earthy	1 Campbell Soup Company fresh packaged mushrooms—diced (Camden, NJ)

Please note that for most of these attributes very precise reference standards were created—this is the ideal. But for the attribute in bold a definition and no reference standard is given—this is not an ideal situation

[a] All solutions made using Culligan sodium-free drinking water (Culligan Water Ltd., Columbia, MO)

[b] All other flavor standards were the same as those for aroma standards

the suitability of each potential reference standard for cheddar cheese by having them score the attributes on an appropriateness scale.

During the final training session the panelists create the score sheet. They may be allowed to decide on the scale to use, although in our laboratories we usually use either the unstructured line scale (similar to Fig. 10.1) or the 15-point unlabeled box scale (Fig. 10.4) for most studies.

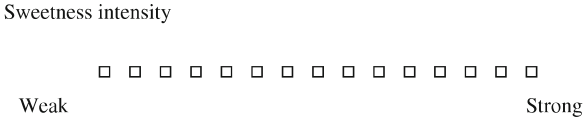

Fig. 10.4 Example of a 15-point unlabeled box scale.

The panelists are asked to decide on the words needed to anchor the scales such as none to extreme or slight to very strong. We also frequently allow the panelists to determine the sequence in which they would like to evaluate the attributes, for example, visual attributes first (unless these are performed separately in a color evaluation chamber such as the Gretag-MacBeth Judge II); then aroma; followed by taste, flavor-by-mouth, and mouth feel; and lastly, after expectoration or swallowing, after-taste. For some panels this order may be changed—for example they may choose to do the taste, flavor by mouth, and mouth feel terms prior to aroma. Once again, the panel leader makes sure that the panelists are comfortable with all the terms, references, and definitions used. At this point the panel leader will start to evaluate judge reproducibility.

A typical sequence of "ballot training" sessions would be the following: Initially, the panelists are exposed to the entire range of the products. They are asked to evaluate the sensory differences among the samples. This occurs in silence. When all panelists complete this portion of the assignment, the panel leader gives each panelist a word list (or sample score sheet) for the products. The word list contains words, definitions, and often the panel leader will also have reference standards available to anchor the descriptors. There are a number of published word lists (lexicons) available for a variety of food and personal care products. A non-exhaustive list is given at the end of this section. The panelists are then asked to indicate through consensus which of these words, reference standards, and definitions should be used in the specific study. The panelists are allowed to add or delete terms through consensus. They are also asked to sequence the descriptors on the ballot.

In subsequent sessions the panelists are exposed to the samples again and asked to look at the ballot that they previously created. They then have to decide if this is truly the score sheet they want to use with these products. Refinement of the score sheet, reference standards, and definitions continues until the panelists are satisfied that this is the best possible score sheet, best sequence, and that everyone understands each term completely. Now the panel leader is ready to determine judge reproducibility.

Some of the available sensory lexicons (vocabularies) are the ASTM publications that cover a large number of product categories (Civille and Lyon, 1996; Rutledge, 2004) as well as Drake and Civille (2003) which covers the creation of flavor lexicons and has numerous references to available word lists. A few recent word lists are Cliff et al. (2000) for apple juices, Dooley et al. (2009) for lip products, Drake et al. (2007) for soy and whey proteins in two countries, Retiveau et al. (2005) for French cheeses, Lee and Chambers (2007) for green tea, Krinsky et al. (2006) for edamame beans, and Riu-Aumatell et al. (2008) for dry gins. There are also published reports of generic descriptive analysis using terminology that are extremely localized. An example would be Nindjin et al. (2007) who trained a group of adult villagers in the Ivory Coast to use the local language to describe the sensory differences among samples of "foutou" (pounded yams).

10.5.1.2 Determining Panelist Reproducibility During Training

Immediately after the training phase the panelists are told that the evaluation phase of the study will begin. However, in reality, the first two or three sessions are used to determine judge consistency. A subset of samples to be used for the real study is served to the panelists in triplicate. The data from these sessions are analyzed; the sensory scientist will study the significance levels of the interaction effects associated with panelists. In a well-trained panel these effects would be not significantly different among judges. If there are significant panelist-associated interaction effects the sensory scientist will determine which judges should be further trained in the use of which descriptors. If all judges are not reproducible then they all need to return to the training phase. However, the results usually indicate that one or two subjects have problems with one or two descriptors. These problems can usually be resolved during a few one-on-one training sessions. Cliff et al. (2000) showed that as training progressed the standard deviations associated with 10 of their 16 attributes decreased. In some cases this decrease was large (0.90 on a 10 cm line scale for oxidized aroma and flavor) and in others smaller (<0.05 for green-grassy and sour). Their panelists anecdotally found that the biggest training effects occurred when the chosen reference standards were unambiguous. See below for a more in-depth discussion on panel performance monitoring.

Recently some work on the effect of feedback calibration on panel training has been published (Findlay et al., 2006, 2007). These authors found that immediate graphical computerized feedback on performance in the sensory booths during training led to reduced training time as well as excellent panel performance. McDonell et al. (2001) also found that feedback in the form of principal component analysis plots, analysis of variance shown to the panel after each descriptive analysis sped up the training process and made the panel more consistent. Nogueira-Terrones et al. (2008) trained a descriptive panel over the Internet to evaluate sausages. Their training process essentially involved feedback on performance at each session and increased training duration increased their Internet panelists' performance relative to the performance of panelists trained more conventionally. However, Marchisano et al. (2000) had found that feedback was positive

for recognition tests, had no effect on discrimination tests (triangle tests), and may have been a negative for scaling tests. Clearly, additional studies are needed. There is an ongoing discussion in sensory circles as to whether panelists should be recruited from within or from outside companies, in other words, whether company employees should be expected to volunteer for panel duty as part of their other duties or whether panelists should be employed to only be on sensory panels. There is very little research to guide on in this discussion. One of the few studies was the one by Lund et al. (2009). They surveyed panelists in New Zealand, Australia, Spain, and the United States and found that the key drivers stimulating people to participate in sensory panels were a general interest in food and extra income. Additionally, panelists on external panels (those not otherwise employed by the company) were more intrinsically motivated than internal panelists (those otherwise employed by the company). Panelists' experience also improved their intrinsic motivation.

10.5.1.3 Evaluating Samples

Standard sensory practices, such as sample coding, randomized serving sequences, use of individual booths, should be employed during the evaluation phase of the study. The sample preparation and serving should also be standardized. The judges should evaluate all samples in at least duplicate, but preferably in triplicate. Samples are usually served monadically and all attributes for a specific sample are evaluated before the next sample is served. However, as shown by Mazzucchelli and Guinard (1999) and Hein (2005) there are no major differences between the results when samples are served monadically or simultaneously (all samples served together and attributes rated one at a time across samples). However, in both studies the actual time taken to do the evaluation increased for the simultaneous serving condition. Under ideal conditions, all samples will be served in a single session, with different sessions as the replicates. If it is not possible to do so then an appropriate experimental plan such as a Latin square, balanced incomplete block should be followed (Cochran and Cox, 1957; Petersen, 1985). The data are usually analyzed by analysis of variance. However, analysis by one or more appropriate multivariate statistical techniques may yield additional information (see Chapter 18).

10.5.1.4 Panel Performance Monitoring

As stated in Section 10.5.1.2—Determining Panelist reproducibility during training—, the sensory scientist will usually have panelists evaluate a subset of products in replicate and then analyze that data to determine whether further training is warranted. However, one may also be interested in monitoring panelist performance over the life span of the panel. This is more usually done when a panel continues to be used over a number of projects or for a number of years, i.e., when one has a "permanent panel." For example, some of the panelists in the Kansas State University Sensory Analysis Center panel have been participating in the panel since 1982 (personal communication, Edgar Chambers, IV, October 2009). When one has a "temporary panel"—a panel that is trained for one specific project and then disbanded—it is more unusual to do ongoing panelist performance monitoring. One may also be interested in panelist performance monitoring when newly trained panelists are merged into an ongoing panel, something that occurs routinely in many commercial settings.

The techniques used to monitor panel performance are similar whether one is monitoring the panel toward the end of training or for the other reasons listed above. The key pieces of information the sensory scientist needs are (a) individual panelist discriminating ability; (b) individual panelist reproducibility; (c) individual panelist agreement with the panel as a whole; (d) panel discriminating ability; and (e) panel reproducibility. Numerous statistical analyses are available to find these pieces of information from the panel data. Please see Meullenet et al. (2007), Tomic et al. (2007), and Martin and Lengard (2005) for additional information on this topic. Derndorfer and coworkers (2005) published code in R to evaluate panel performance. Pineau et al. (2007) published a mixed-model and control chart approach using SAS (SI, Cary, NC). SensomineR (a freeware R-package) also contains panel performance techniques, as well as many sensory data analysis techniques (Lê and Husson, 2008). Additionally, Panel Check, another freeware R-based program, is available for download at http://www.panelcheck.com/ (Kollár-Hunek et al., 2007; Tomic et al., 2007). In this

section we will briefly discuss four of these techniques. In order to simplify the discussion of panel performance monitoring we assume that each member of the sensory panel evaluated the entire set of products in triplicate.

Univariate Techniques

One-way analyses of variance with product as the main effect for each panelist and each attribute allow the sensory scientist to evaluate the individual panelists' discriminability as well as their repeatability. The assumption is that panelists with excellent discriminability for a specific attribute would have large F-values and small probability (p) values. Panelists with good repeatability would tend to have small mean square error (MSE) values. A plot of p-values by MSE values allows the sensory scientist to simultaneously evaluate both discriminability and repeatability (Fig. 10.5).

A three-way analysis of variance with main effects (product, panelist, and replication) and interaction effects (panelist by product, panelist by replication, and product by replication) will fairly quickly indicate some trouble spots in panelist performance relative to the rest of the panel. The sensory scientist should be on the lookout for attributes with significant panelist by product interactions. These would indicate that at least one (and possibly more panelists) is not scoring these attributes similarly. One should always plot the data. If a panelist's results decrease (increase) while the panel means increase (decrease) it is called a cross-over interaction and it is a major problem. If a panelist's results decrease (increase) while the panel means decrease (increase) but at a different rate then the interaction is less of a problem.

Panelist performance relative to the panel as a whole for each attribute can also be visually shown with eggshell plots (Hirst and Næs, 1994). In this case the panelist's scores for each attribute are transformed into ranks. A consensus ranking for each attribute is then created by finding the mean rank over panelists for each product and then ranking these means. Each panelist's cumulative scores are then plotted relative to the consensus ranks. The resultant plot looks similar to an eggshell, and the intention is to have as few "cracks" as possible in the shell for each attribute (Fig. 10.6).

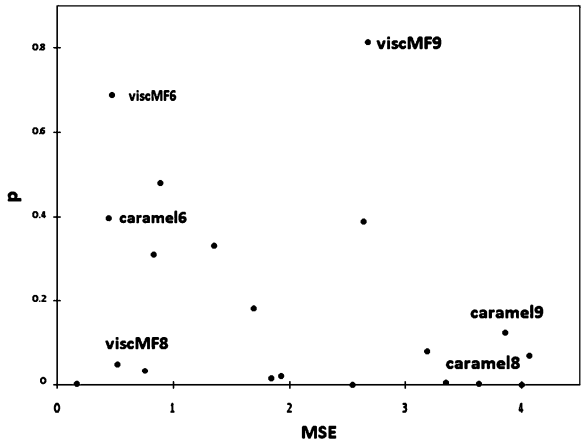

Fig. 10.5 An example of a p by MSE plot for all panelists (only some panelists are named) for caramel aroma (caramel) and viscous mouth feel (viscMF). Panelist 8 shows excellent discriminability (low p-value) as well as excellent repeatability (low MSE) for viscous mouth feel. For caramel this panelist also has excellent discriminability (low p-value) despite repeatability issues. Panelist 9 has repeatability issues, especially for both attributes but also has discriminability issues with viscous mouth feel and to a lesser extent with caramel aroma.

Multivariate Techniques

A principal component analysis (PCA) of each attribute for all the panelists will indicate the consonance (agreement) among the panelists for that attribute (Dijksterhuis, 1995). In this case the panelist scores for each product for the specified attribute are used as the variables (columns) in the analysis. If there is substantial agreement (consonance) among the panelists then the majority of the variance should be explained by the first dimension. In other words if the panelists use the specific attribute similarly then the PCA should tend to become unidimensional. Usually, for well-trained panels the amount of variance explained on the first dimension ranges from about 50 to 70% (Fig. 10.7).

Worch et al. (2009) found that for untrained consumers these values tend to be much lower, ranging from about 15 to about 24%. The sensory scientist can also calculate a consonance (C) score for each attribute from the PCA results. Dijksterhuis (1995) defined C as the ratio of the variance explained by

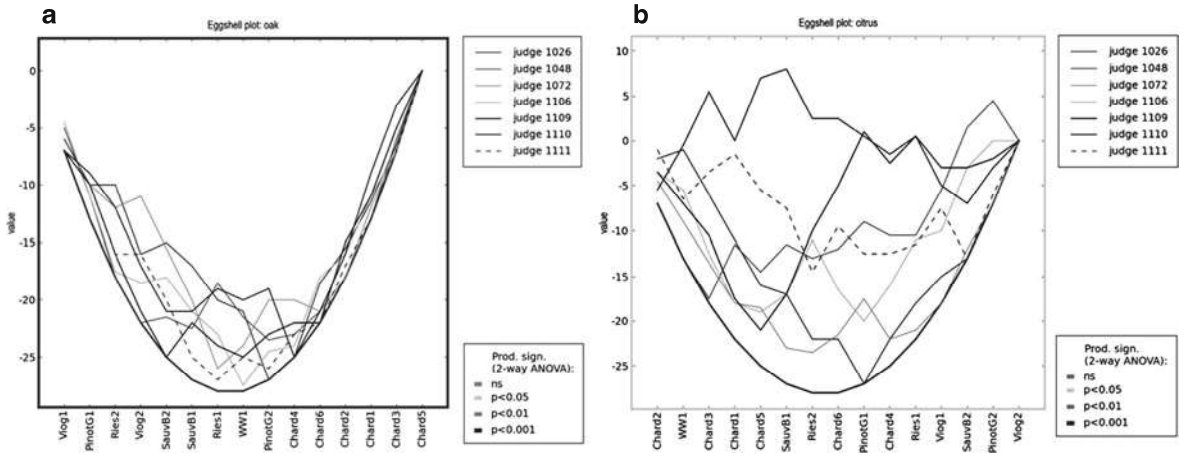

Fig. 10.6 Two examples of eggshell plots. The smooth line on the bottom of the plot is the consensus rank for the specific attribute. From the plots it is evident that the panelists were more in agreement with each other on the oak aroma attribute (**a**) than on the citrus aroma attribute (**b**).

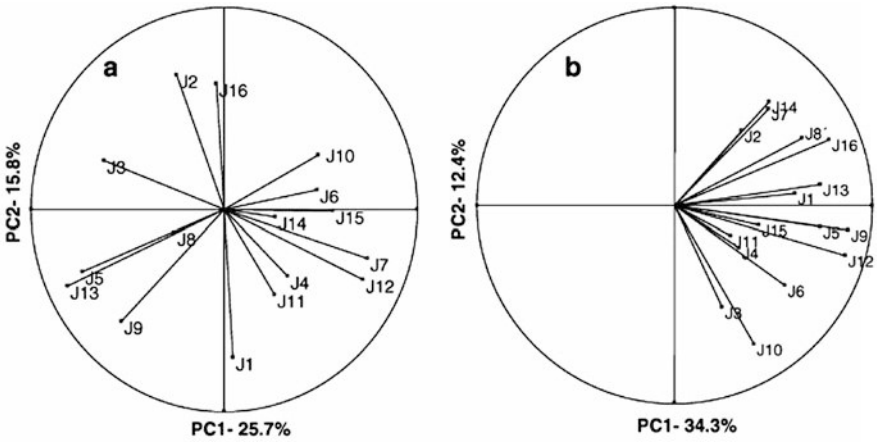

Fig. 10.7 An example of two PCA panelist consonance plots. In the first plot (**a**) there is disagreement among the panelists in their usage of the specific term. In the second plot (**b**) there is more agreement among the panelists in terms of their use of the second term (reprinted with permission from Le Moigne et al., 2008).

the first dimension to the sum of the remaining variances. Large values of C would indicate that there was agreement among the panelists in the usage of a specified term since the vectors for the terms would "point" in the same direction. The sensory scientist must be careful to not just blindly calculate C since large values of C are possible when there are large negative loadings on the first dimension as well as large positive ones. Thus prior to calculating C one should always plot the PCA for each attribute. Dellaglio et al. (1996) reported C values ranging from about 0.4 to 2.3 for a panel evaluating Italian dry-cured sausages. Carbonell et al. (2007) found C values ranging from 0.46 to 4.6 for a panel evaluating Spanish mandarin juices.

10.5.2 Studies Comparing Different Conventional Descriptive Analysis Techniques

Risvik et al. (1992) and Heymann (1994a) found that well-trained independent panels (in two different countries, Norway and Britain, and in the same university setting, respectively) gave very comparable results. A study by Lotong et al. (2002) on the evaluation of orange juices by two independently highly trained panels (one panel used individual judgments and the other created a consensus evaluation) showed that the results from the different panels were comparable. Drake et al. (2007) evaluated the descriptive sensory analyses of