Chapter 1

Introduction

Abstract In this chapter we carefully parse the definition for sensory evaluation, briefly discuss validity of the data collected before outlining the early history of the field. We then describe the three main methods used in sensory evaluation (discrimination tests, descriptive analysis, and hedonic testing) before discussing the differences between analytical and consumer testing. We then briefly discuss why one may want to collect sensory data. In the final sections we highlight the differences and similarities between sensory evaluation and marketing research and between sensory evaluation and commodity grading as used in, for example, the dairy industry.

Sensory evaluation is a child of industry. It was spawned in the late 40's by the rapid growth of the consumer product companies, mainly food companies.... Future development in sensory evaluation will depend upon several factors, one of the most important being the people and their preparation and training.

— Elaine Skinner (1989)

Contents

1.1 Introduction and Overview

1.1.1 Definition

The field of sensory evaluation grew rapidly in the second half of the twentieth century, along with the expansion of the processed food and consumer products industries. Sensory evaluation comprises a set of techniques for accurate measurement of human responses to foods and minimizes the potentially biasing effects of brand identity and other information influences on consumer perception. As such, it attempts to isolate the sensory properties of foods themselves and provides important and useful information to product developers, food scientists, and managers about the sensory characteristics of their products. The field was comprehensively reviewed by Amerine, Pangborn, and Roessler in 1965, and more recent texts have been published by Moskowitz et al. (2006), Stone and Sidel (2004), and Meilgaard et al. (2006). These three later sources are practical works aimed at sensory specialists

H.T. Lawless, H. Heymann, *Sensory Evaluation of Food*, Food Science Text Series,
DOI 10.1007/978-1-4419-6488-5_1, © Springer Science+Business Media, LLC 2010

in industry and reflect the philosophies of the consulting groups of the authors. Our goal in this book is to provide a comprehensive overview of the field with a balanced view based on research findings and one that is suited to students and practitioners alike.

Sensory evaluation has been defined as a scientific method used to evoke, measure, analyze, and interpret those responses to products as perceived through the senses of sight, smell, touch, taste, and hearing (Stone and Sidel, 2004). This definition has been accepted and endorsed by sensory evaluation committees within various professional organizations such as the Institute of Food Technologists and the American Society for Testing and Materials. The principles and practices of sensory evaluation involve each of the four activities mentioned in this definition. Consider the words "to evoke." Sensory evaluation gives guidelines for the preparation and serving of samples under controlled conditions so that biasing factors are minimized. For example, people in a sensory test are often placed in individual test booths so that the judgments they give are their own and do not reflect the opinions of those around them. Samples are labeled with random numbers so that people do not form judgments based upon labels, but rather on their sensory experiences. Another example is in how products may be given in different orders to each participant to help measure and counterbalance for the sequential effects of seeing one product after another. Standard procedures may be established for sample temperature, volume, and spacing in time, as needed to control unwanted variation and improve test precision.

Next, consider the words, "to measure." Sensory evaluation is a quantitative science in which numerical data are collected to establish lawful and specific relationships between product characteristics and human perception. Sensory methods draw heavily from the techniques of behavioral research in observing and quantifying human responses. For example, we can assess the proportion of times people are able to discriminate small product changes or the proportion of a group that expresses a preference for one product over another. Another example is having people generate numerical responses reflecting their perception of how strong a product may taste or smell. Techniques of behavioral research and experimental psychology offer guidelines as to how such measurement techniques should be employed and what their potential pitfalls and liabilities may be.

The third process in sensory evaluation is analysis. Proper analysis of the data is a critical part of sensory testing. Data generated from human observers are often highly variable. There are many sources of variation in human responses that cannot be completely controlled in a sensory test. Examples include the mood and motivation of the participants, their innate physiological sensitivity to sensory stimulation, and their past history and familiarity with similar products. While some screening may occur for these factors, they may be only partially controlled, and panels of humans are by their nature heterogeneous instruments for the generation of data. In order to assess whether the relationships observed between product characteristics and sensory responses are likely to be real, and not merely the result of uncontrolled variation in responses, the methods of statistics are used to analyze evaluation data. Hand-in-hand with using appropriate statistical analyses is the concern of using good experimental design, so that the variables of interest are investigated in a way that allows sensible conclusions to be drawn.

The fourth process in sensory evaluation is the interpretation of results. A sensory evaluation exercise is necessarily an experiment. In experiments, data and statistical information are only useful when interpreted in the context of hypotheses, background knowledge, and implications for decisions and actions to be taken. Conclusions must be drawn that are reasoned judgments based upon data, analyses, and results. Conclusions involve consideration of the method, the limitations of the experiment, and the background and contextual framework of the study. The sensory evaluation specialists become more than mere conduits for experimental results, but must contribute interpretations and suggest reasonable courses of action in light of the numbers. They should be full partners with their clients, the end-users of the test results, in guiding further research. The sensory evaluation professional is in the best situation to realize the appropriate interpretation of test results and the implications for the perception of products by the wider group of consumers to whom the results may be generalized. The sensory specialist best understands the limitations of the test procedure and what its risks and liabilities may be.

A sensory scientist who is prepared for a career in research must be trained in all four of the phases mentioned in the definition. They must understand products, people as measuring instruments, statistical analyses, and interpretation of data within the context of research objectives. As suggested in Skinner's quote, the future advancement of the field depends upon the breadth and depth of training of new sensory scientists.

1.1.2 Measurement

Sensory evaluation is a science of measurement. Like other analytical test procedures, sensory evaluation is concerned with precision, accuracy, sensitivity, and avoiding false positive results (Meiselman, 1993). Precision is similar to the concept in the behavioral sciences of reliability. In any test procedure, we would like to be able to get the same result when a test is repeated. There is usually some error variance around an obtained value, so that upon repeat testing, the value will not always be exactly the same. This is especially true of sensory tests in which human perceptions are necessarily part of the generation of data. However, in many sensory test procedures, it is desirable to minimize this error variance as much as possible and to have tests that are low in error associated with repeated measurements. This is achieved by several means. As noted above, we isolate the sensory response to the factors of interest, minimizing extraneous influences, controlling sample preparation and presentation. Additionally, as necessary, sensory scientists screen and train panel participants.

A second concern is the accuracy of a test. In the physical sciences, this is viewed as the ability of a test instrument to produce a value that is close to the "true" value, as defined by independent measurement from another instrument or set of instruments that have been appropriately calibrated. A related idea in the behavioral sciences, this principle is called the validity of a test. This concerns the ability of a test procedure to measure what it was designed and intended to measure. Validity is established in a number of ways. One useful criterion is predictive validity, when a test result is of value in predicting what would occur in another situation or another measurement. In sensory testing, for example, the test results should reflect the perceptions and opinions of consumers that might buy the product. In other words, the results of the sensory test should generalize to the larger population. The test results might correlate with instrumental measures, process or ingredient variables, storage factors, shelf life times, or other conditions known to affect sensory properties. In considering validity, we have to look at the end use of the information provided by a test. A sensory test method might be valid for some purposes, but not others (Meiselman, 1993). A simple difference test can tell if a product has changed, but not whether people will like the new version.

A good sensory test will minimize errors in measurement and errors in conclusions and decisions. There are different types of errors that may occur in any test procedure. Whether the test result reflects the true state of the world is an important question, especially when error and uncontrolled variability are inherent in the measurement process. Of primary concern in sensory tests is the sensitivity of the test to differences among products. Another way to phrase this is that a test should not often miss important differences that are present. "Missing a difference" implies an insensitive test procedure. To keep sensitivity high, we must minimize error variance wherever possible by careful experimental controls and by selection and training of panelists where appropriate. The test must involve sufficient numbers of measurements to insure a tight and reliable statistical estimate of the values we obtain, such as means or proportions. In statistical language, detecting true differences is avoiding Type II error and the minimization of β-risk. Discussion of the power and sensitivity of tests from a statistical perspective occurs in Chapter 5 and in the Appendix.

The other error that may occur in a test result is that of finding a positive result when none is actually present in the larger population of people and products outside the sensory test. Once again, a positive result usually means detection of a statistically significant difference between test products. It is important to use a test method that avoids false positive results or Type I error in statistical language. Basic statistical training and common statistical tests applied to scientific findings are oriented toward avoiding this kind of error. The effects of random chance deviations must be taken into account in deciding if a test result reflects a real difference or whether our result is likely to be due to chance variation. The common procedures of inferential statistics provide assurance that we have limited our possibility of finding a difference where one does not really exist. Statistical procedures reduce this risk to some comfortable level, usually with a ceiling of 5% of all tests we conduct.

Note that this error of a false positive experimental result is potentially devastating in basic scientific research—whole theories and research programs may develop from spurious experimental implications if results are due to only random chance. Hence we guard against this kind of danger with proper application of statistical tests. However, in product development work, the second kind of statistical error, that of missing a true difference can be equally devastating. It may be that an important ingredient or processing change has made the product better or worse from a sensory point of view, and this change has gone undetected. So sensory testing is equally concerned with not missing true differences and with avoiding false positive results. This places additional statistical burdens on the experimental concerns of sensory specialists, greater than those in many other branches of scientific research.

Finally, most sensory testing is performed in an industrial setting where business concerns and strategic decisions enter the picture. We can view the outcome of sensory testing as a way to reduce risk and uncertainty in decision making. When a product development manager asks for a sensory test, it is usually because there is some uncertainty about exactly how people perceive the product. In order to know whether it is different or equivalent to some standard product, or whether it is preferred to some competitive standard, or whether it has certain desirable attributes, data are needed to answer the question. With data in hand, the end-user can make informed choices under conditions of lower uncertainty or business risk. In most applications, sensory tests function as risk reduction mechanisms for researchers and marketing managers.

In addition to the obvious uses in product development, sensory evaluation may provide information to other corporate departments. Packaging functionality and convenience may require product tests. Sensory criteria for product quality may become an integral part of a quality control program. Results from blind-labeled sensory consumer tests may need to be compared to concept-related marketing research results. Sensory groups may even interact with corporate legal departments over advertising claim substantiation and challenges to claims. Sensory evaluation also functions in situations outside corporate research. Academic research on foods and materials and their properties and processing will often require sensory tests to evaluate the human perception of changes in the products (Lawless and Klein, 1989). An important function of sensory scientists in an academic setting is to provide consulting and resources to insure that quality tests are conducted by other researchers and students who seek to understand the sensory impact of the variables they are studying. In government services such as food inspection, sensory evaluation plays a key role (York, 1995). Sensory principles and appropriate training can be key in insuring that test methods reflect the current knowledge of sensory function and test design. See Lawless (1993) for an overview of the education and training of sensory scientists—much of this piece still rings true more than 15 years later.

1.2 Historical Landmarks and the Three Classes of Test Methods

The human senses have been used for centuries to evaluate the quality of foods. We all form judgments about foods whenever we eat or drink ("Everyone carries his own inch-rule of taste, and amuses himself by applying it, triumphantly, wherever he travels."—Henry Adams, 1918). This does not mean that all judgments are useful or that anyone is qualified to participate in a sensory test. In the past, production of good quality foods often depended upon the sensory acuity of a single expert who was in charge of production or made decisions about process changes in order to make sure the product would have desirable characteristics. This was the historical tradition of brewmasters, wine tasters, dairy judges, and other food inspectors who acted as the arbiters of quality. Modern sensory evaluation replaced these single authorities with panels of people participating in specific test methods that took the form of planned experiments. This change occurred for several reasons. First, it was recognized that the judgments of a panel would in general be more reliable than the judgments of single individual and it entailed less risk since the single expert could become ill, travel, retire, die, or be otherwise unavailable to make decisions. Replacement of such an individual was a nontrivial problem. Second, the expert might or might not reflect what consumers or segments of the consuming public might want in a product. Thus for issues of product quality and overall appeal, it was safer (although often more time consuming and expensive) to go directly to the target population. Although the tradition of informal, qualitative inspections such as benchtop “cuttings” persists in some industries, they have been gradually replaced by more formal, quantitative, and controlled observations (Stone and Sidel, 2004).

The current sensory evaluation methods comprise a set of measurement techniques with established track records of use in industry and academic research. Much of what we consider standard procedures comes from pitfalls and problems encountered in the practical experience of sensory specialists over the last 70 years of food and consumer product research, and this experience is considerable. The primary concern of any sensory evaluation specialist is to insure that the test method is appropriate to answer the questions being asked about the product in the test. For this reason, tests are usually classified according to their primary purpose and most valid use. Three types of sensory testing are commonly used, each with a different goal and each using participants selected using different criteria. A summary of the three main types of testing is given in Table 1.1.

1.2.1 Difference Testing

The simplest sensory tests merely attempt to answer whether any perceptible difference exists between two types of products. These are the discrimination tests or simple difference testing procedures. Analysis is usually based on the statistics of frequencies and proportions (counting right and wrong answers). From the test results, we infer differences based on the proportions of persons who are able to choose a test product correctly from among a set of similar or control products. A classic example of this test was the triangle procedure, used in the Carlsberg breweries and in the Seagrams distilleries in the 1940s (Helm and Trolle, 1946; Peryam and Swartz, 1950). In this test, two products were from the same batch while a third product was different. Judges would be asked to pick the odd sample from among the three. Ability to discriminate differences would be inferred from consistent correct choices above the level expected by chance. In breweries, this test served primarily as a means to screen judges for beer evaluation, to insure that they possessed sufficient discrimination abilities. Another multiple-choice difference test was developed at about the same time in distilleries for purposes of quality control (Peryam and Swartz, 1950). In the duo–trio procedure, a reference sample was given and then two test samples. One of the test samples matched the reference while the other was from a different product, batch or process. The participant would try to match the correct sample to the reference, with a chance probability of one-half. As in the triangle test, a proportion of correct choices above that expected by chance is considered evidence for a perceivable difference between products. A third popular difference test was the paired comparison, in which participants would be asked to choose which of two products was stronger or more intense in a given attribute. Partly due to the fact that the panelist’s attention is directed to a specific attribute, this test is very sensitive to differences. These three common difference tests are shown in Fig. 1.1.

Simple difference tests have proven very useful in application and are in widespread use today. Typically a discrimination test will be conducted with 25–40 participants who have been screened for their sensory acuity to common product differences and who are familiar with the test procedures. This generally provides an adequate sample size for documenting clear sensory differences. Often a replicate test is performed while the respondents are present in the sensory test facility. In part, the popularity of these tests is due to the simplicity of data analysis. Statistical tables derived from the binomial distribution give the minimum number of correct responses needed to conclude statistical

Table 1.1 Classification of test methods in sensory evaluation

Class	Question of interest	Type of test	Panelist characteristics
Discrimination	Are products perceptibly different in any way	“Analytic”	Screened for sensory acuity, oriented to test method, sometimes trained
Descriptive	How do products differ in specific sensory characteristics	“Analytic”	Screened for sensory acuity and motivation, trained or highly trained
Affective	How well are products liked or which products are preferred	“Hedonic”	Screened for products, untrained

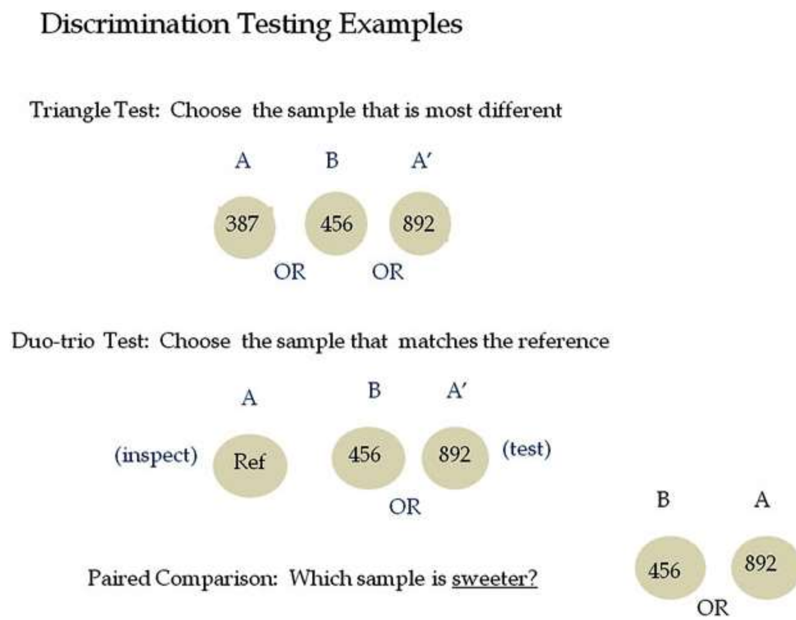

Fig. 1.1 Common methods for discrimination testing include the triangle, duo–trio, and paired comparison procedures.

significance as a function of the number of participants. Thus a sensory technician merely needs to count answers and refer to a table to give a simple statistical conclusion, and results can be easily and quickly reported.

1.2.2 Descriptive Analyses

The second major class of sensory test methods is those that quantify the perceived intensities of the sensory characteristics of a product. These procedures are known as descriptive analyses. The first method to do this with a panel of trained judges was the Flavor Profile® method developed at the Arthur D. Little consulting group in the late 1940s (Caul, 1957). This group was faced with developing a comprehensive and flexible tool for analysis of flavor to solve problems involving unpleasant off flavors in nutritional capsules and questions about the sensory impact of monosodium glutamate in various processed foods. They formulated a method involving extensive training of panelists that enabled them to characterize all of the flavor notes in a food and the intensities of these notes using a simple category scale and noting their order of appearance. This advance was noteworthy on several grounds. It supplanted the reliance on single expert judges (brewmasters, coffee tasters, and such) with a panel of individuals, under the realization that the consensus of a panel was likely to be more reliable and accurate than the judgment of a single individual. Second, it provided a means to characterize the individual attributes of flavor and provide a comprehensive analytical description of differences among a group of products under development.

Several variations and refinements in descriptive analysis techniques were forthcoming. A group at the General Foods Technical Center in the early 1960s developed and refined a method to quantify food texture, much as the flavor profile had enabled the quantification of flavor properties (Brandt et al., 1963, Szczesniak et al., 1975). This technique, the Texture Profile method, used a fixed set of force-related and shape-related attributes to characterize the rheological and tactile properties of foods and how these changed over time with mastication. These characteristics have parallels in the physical evaluation of food breakdown or flow. For example, perceived hardness is related to the physical force required to penetrate a sample. Perceived thickness of a fluid or semisolid is related in part to physical viscosity. Texture profile panelists were also trained to recognize specific intensity points along each scale, using standard products or formulated pseudo-foods for calibration.

Other approaches were developed for descriptive analysis problems. At Stanford Research Institute in the early 1970s, a group proposed a method for descriptive analysis that would remedy some of the apparent shortcomings of the Flavor Profile® method and be even more broadly applicable to all sensory properties of a food, and not just taste and texture (Stone et al., 1974). This method was termed Quantitative Descriptive Analysis® or QDA® for short (Stone and Sidel, 2004). QDA® procedures borrowed heavily from the traditions of behavioral research and used experimental designs and statistical analyses such as analysis of variance. This insured independent judgments of panelists and statistical testing, in contrast to the group discussion and consensus procedures of the Flavor Profile® method. Other variations on descriptive procedures were tried and achieved some popularity, such as the Spectrum Method® (Meilgaard et al., 2006) that included a high degree of calibration of panelists for intensity scale points, much like the Texture Profile. Still other researchers have employed hybrid techniques that include some features of the various descriptive approaches (Einstein, 1991). Today many product development groups use hybrid approaches as the advantages of each may apply to the products and resources of a particular company.

Descriptive analysis has proven to be the most comprehensive and informative sensory evaluation tool. It is applicable to the characterization of a wide variety of product changes and research questions in food product development. The information can be related to consumer acceptance information and to instrumental measures by means of statistical techniques such as regression and correlation.

An example of a descriptive ballot for texture assessment of a cookie product is shown in Table 1.2. The product is assessed at different time intervals in a uniform and controlled manner, typical of an analytical sensory test procedure. For example, the first bite may be defined as cutting with the incisors. The panel for such an analysis would consist of perhaps 10–12 well-trained individuals, who were oriented to the meanings of the terms and given practice with examples. Intensity references to exemplify scale points are also given in some techniques. Note the amount of detailed information that can be provided in this example and bear in mind that this is only looking at the product's texture—flavor might form an equally detailed sensory analysis, perhaps with a separate trained panel. The relatively small number of panelists (a dozen or so) is justified due to their level of calibration. Since they have been trained to use attribute scales in a similar manner, error variance is lowered and statistical power and test sensitivity are maintained in spite of fewer observations (fewer data points per product). Similar examples of texture, flavor, fragrance, and tactile analyses can be found in Meilgaard et al. (2006).

Table 1.2 Descriptive evaluation of cookies–texture attributes

Phase	Attributes	Word anchors
Surface	Roughness	Smooth–rough
	Particles	None–many
	Dryness	Oily–dry
First bite	Fracturability	Crumbly–brittle
	Hardness	Soft–hard
	Particle size	Small–large
First chew	Denseness	Airy–dense
	Uniformity of chew	Even–uneven
Chew down	Moisture absorption	None–much
	Cohesiveness of mass	Loose–cohesive
	Toothpacking	None–much
	Grittiness	None–much
Residual	Oiliness	Dry–oily
	Particles	None–many
	Chalky	Not chalky–very chalky

1.2.3 Affective Testing

The third major class of sensory tests is those that attempt to quantify the degree of liking or disliking of a product, called hedonic or affective test methods. The most straightforward approach to this problem is to offer people a choice among alternative products and see if there is a clear preference from the majority of respondents. The problem with choice tests is that they are not very informative about the magnitude of liking or disliking from respondents. An historical landmark in this class of tests was the hedonic scale developed at the U.S. Army Food and Container Institute in the late 1940s (Jones et al., 1955). This method provided a balanced 9-point scale for liking with a centered neutral category and attempted to produce scale point labels with adverbs that represented psychologically equal steps or changes in hedonic tone. In other words, it was a scale with ruler-like properties whose equal intervals would be amenable to statistical analysis.

An example of the 9-point scale is shown in Fig. 1.2. Typically a hedonic test today would involve a sample of 75–150 consumers who were regular users of the product. The test would involve several alternative versions of the product and be conducted in some central location or sensory test facility. The larger panel size

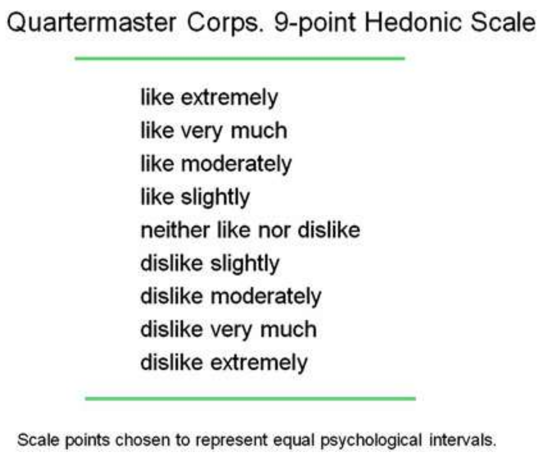

Fig. 1.2 The 9-point hedonic scale used to assess liking and disliking. This scale, originally developed at the U.S. Army Food and Container Institute (Quartermaster Corps), has achieved widespread use in consumer testing of foods.

of an affective test arises due to the high variability of individual preferences and thus a need to compensate with increased numbers of people to insure statistical power and test sensitivity. This also provides an opportunity to look for segments of people who may like different styles of a product, for example, different colors or flavors. It may also provide an opportunity to probe for diagnostic information concerning the reasons for liking or disliking a product.

Workers in the food industry were occasionally in contact with psychologists who studied the senses and had developed techniques for assessing sensory function (Moskowitz, 1983). The development of the 9-point hedonic scale serves as good illustration of what can be realized when there is interaction between experimental psychologists and food scientists. A psychological measurement technique called Thurstonian scaling (see Chapter 5) was used to validate the adverbs for the labels on the 9-point hedonic scale. This interaction is also visible in the authorship of this book—one author is trained in food science and chemistry while the other is an experimental psychologist. It should not surprise us that interactions would occur and perhaps the only puzzle is why the interchanges were not more sustained and productive. Differences in language, goals, and experimental focus probably contributed to some difficulties. Psychologists were focused primarily on the individual person while sensory evaluation specialists were focused primarily on the food product (the stimulus). However, since a sensory perception involves the necessary interaction of a person with a stimulus, it should be apparent that similar test methods are necessary to characterize this person–product interaction.

1.2.4 The Central Dogma—Analytic Versus Hedonic Tests

The central principle for all sensory evaluation is that the test method should be matched to the objectives of the test. Figure 1.3 shows how the selection of the test procedure flows from questions about the objective of the investigation. To fulfill this goal, it is necessary to have clear communication between the sensory test manager and the client or end-user of the information. A dialogue is often needed. Is the important question whether or not there is any difference at all among the products? If so, a discrimination test is indicated. Is the question one of whether consumers like the new product better than the previous version? A consumer acceptance test is needed. Do we need to know what attributes have changed in the sensory characteristics of the new product? Then a descriptive analysis procedure is called for. Sometimes there are multiple objectives and a sequence of different tests is required (Lawless and Claassen, 1993). This can present problems if all the answers are required at once or under severe time pressure during competitive product development. One of the most important jobs of the sensory specialist in the food industry is to insure a clear understanding and specification of the type of information needed by the end-users. Test design may require a number of conversations, interviews with different people, or even written test requests that specify why the information is to be collected and how the results will be used in making specific decisions and subsequent actions to be taken. The sensory specialist is the best position to understand the uses and limitations of each procedure and what would be considered appropriate versus inappropriate conclusions from the data.

There are two important corollaries to this principle. The sensory test design involves not only the selection of an appropriate method but also the selection of appropriate participants and statistical analyses. The three classes of sensory tests can be divided into two types, analytical sensory tests including discrimination

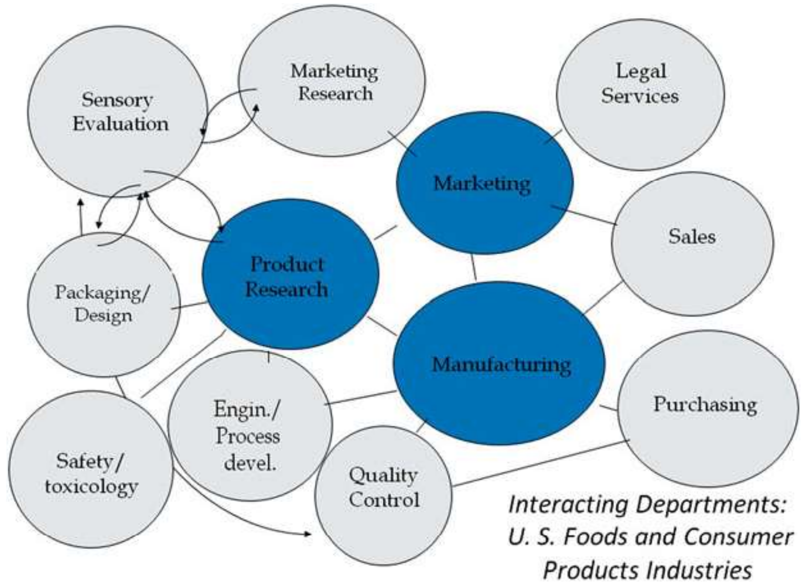

Fig. 1.3 A flowchart showing methods determination. Based on the major objectives and research questions, different sensory test methods are selected. Similar decision processes are made in panelist selection, setting up response scales, in choosing experimental designs, statistical analysis, and other tasks in designing a sensory test (reprinted with permission from Lawless, 1993).

and descriptive methods and affective or hedonic tests such as those involved in assessing consumer liking or preferences (Lawless and Claassen, 1993). For the analytical tests, panelists are selected based on having average to good sensory acuity for the critical characteristics (tastes, smells, textures, etc.) of products to be evaluated. They are familiarized with the test procedures and may undergo greater or lesser amounts of training, depending upon the method. In the case of descriptive analysis, they adopt an analytical frame of mind, focusing on specific aspects of the product as directed by the scales on their questionnaires. They are asked to put personal preferences and hedonic reactions aside, as their job is only to specify what attributes are present in the product and at what levels of sensory intensity, extent, amount, or duration.

In contrast to this analytical frame of mind, consumers in an affective test act in a much more integrative fashion. They perceive a product as a whole pattern. Although their attention is sometimes captured by a specific aspect of a product (especially if it is a bad, unexpected, or unpleasant one), their reactions to the product are often immediate and based on the integrated pattern of sensory stimulation from the product and expressed as liking or disliking. This occurs without a great deal of thought or dissection of the product's specific profile. In other words, consumers are effective at rendering impressions based on the integrated pattern of perceptions. In such consumer tests, participants must be chosen carefully to insure that the results will generalize to the population of interest. Participants should be frequent users of the product, since they are most likely to form the target market and will be familiar with similar products. They possess reasonable expectations and a frame of reference within which they can form an opinion relative to other similar products they have tried.

The analytic/hedonic distinction gives rise to some highly important rules of thumb and some warnings about matching test methods and respondents. It is unwise to ask trained panelists about their preferences or whether they like or dislike a product. They have been asked to assume a different, more analytical frame of mind and to place personal preference aside. Furthermore, they have not necessarily been selected to be frequent users of the product, so they are not part of the target population to which one would like to generalize hedonic test results. A common analogy here is to an analytical instrument. You would not ask a gas chromatograph or a pH meter whether it liked the product, so why ask your analytical descriptive panel (O'Mahony, 1979).

Conversely, problems arise when consumers are asked to furnish very specific information about product attributes. Consumers not only act in a non-analytic frame of mind but also often have very fuzzy concepts about specific attributes, confusing sour and bitter tastes, for example. Individuals often differ markedly in their interpretations of sensory attribute words on a questionnaire. While a trained texture profile panel has no trouble in agreeing how cohesive a product is after chewing, we cannot expect consumers to provide precise information on such a specific and technical attribute. In summary, we avoid using trained panelists for affective information and we avoid asking consumers about specific analytical attributes.

Related to the analytic–hedonic distinction is the question of whether experimental control and precision are to be maximized or whether validity and generalizability to the real world are more important. Often there is a tradeoff between the two and it is difficult to maximize both simultaneously. Analytic tests in the lab with specially screened and trained judges are more reliable and lower in random error than consumer tests. However, we give up a certain amount of generalizability to real-world results by using artificial conditions and a special group of participants. Conversely, in the testing of products by consumers in their own homes we have not only a lot of real-life validity but also a lot of noise in the data. Brinberg and McGrath (1985) have termed this struggle between precision and validity one of "conflicting desiderata." O'Mahony (1988) has made a distinction between sensory evaluation Type I and Type II. In Type I sensory evaluation, reliability and sensitivity are key factors, and the participant is viewed much like an analytical instrument used to detect and measure changes in a food product. In Type II sensory evaluation, participants are chosen to be representative of the consuming population, and they may evaluate food under more naturalistic conditions. Their emphasis here is on prediction of consumer response. Every sensory test falls somewhere along a continuum where reliability versus real-life extrapolation are in a potential tradeoff relationship. This factor must also be discussed with end-users of the data to see where their emphasis lies and what level of tradeoff they find comfortable.

Statistical analyses must also be chosen with an eye to the nature of the data. Discrimination tests involve choices and counting numbers of correct responses. The statistics derived from the binomial distribution or those designed for proportions such as chi-square are appropriate. Conversely, for most scaled data, we can apply the familiar parametric statistics appropriate to normally distributed and continuous data, such as means, standard deviations, *t*-tests, analysis of variance. The choice of an appropriate statistical test is not always straightforward, so sensory specialists are wise to have thorough training in statistics and to involve statistical and design specialists in a complex project in its earliest stages of planning.

Occasionally, these central principles are violated. They should not be put aside as a matter of mere expediency or cost savings and never without a logical analysis. One common example is the use of a discrimination test before consumer acceptance. Although our ultimate interest may lie in whether consumers will like or dislike a new product variation, we can conduct a simple difference test to see whether any change is perceivable at all. The logic in this sequence is the following: if a screened and experienced discrimination panel cannot tell the difference under carefully controlled conditions in the sensory laboratory, then a more heterogeneous group of consumers is unlikely to see a difference in their less controlled and more variable world. If no difference is perceived, there can logically be no systematic preference. So a more time consuming and costly consumer test can sometimes be avoided by conducting a simpler but more sensitive discrimination test first. The added reliability of the controlled discrimination test provides a "safety net" for conclusions about consumer perception. Of course, this logic is not without its pitfalls—some consumers may interact extensively with the product during a home use test period and may form stable and important opinions that are not captured in a short duration laboratory test, and there is also always the possibility of a false negative result (the error of missing a difference). MacRae and Geelhoed (1992) describe an interesting case of a missed difference in a triangle test where a significant preference was then observed between water samples in a paired comparison. The sensory professional must be aware that these anomalies in experimental results will sometimes arise, and must also be aware of some of the reasons why they occur.

1.3 Applications: Why Collect Sensory Data?

Human perceptions of foods and consumer products are the results of complex sensory and interpretation processes. At this stage in scientific history, perceptions of such multidimensional stimuli as conducted by the parallel processing of the human nervous system are difficult or impossible to predict from instrumental measures. In many cases instruments lack the sensitivity of human sensory systems—smell is a good example. Instruments rarely mimic the mechanical manipulation of foods when tasted nor do they mimic the types of peri-receptor filtering that occur in biological fluids like saliva or mucus that can cause chemical partitioning of flavor materials. Most importantly, instrumental assessments give values that miss an important perceptual process: the interpretation of sensory experience by the human brain prior to responding. The brain lies interposed between sensory input and the generation of responses that form our data. It is a massively parallel-distributed processor and computational engine, capable of rapid feats of pattern recognition. It comes to the sensory evaluation task complete with a personal history and experiential frame of reference. Sensory experience is interpreted, given meaning within the frame of reference, evaluated relative to expectations and can involve integration of multiple simultaneous or sequential inputs. Finally judgments are rendered as our data. Thus there is a "chain of perception" rather than simply stimulus and response (Meilgaard et al., 2006).

Only human sensory data provide the best models for how consumers are likely to perceive and react to food products in real life. We collect, analyze, and interpret sensory data to form predictions about how products have changed during a product development program. In the food and consumer products industries, these changes arise from three important factors: ingredients, processes, and packaging. A fourth consideration is often the way a product ages, in other words its shelf life, but we may consider shelf stability to be one special case of processing, albeit usually a very passive one (but also consider products exposed to temperature fluctuation, light-catalyzed oxidation, microbial contamination, and other "abuses"). Ingredient changes arise for a number of reasons. They may be introduced to improve product quality, to reduce costs of production, or simply because a certain supply of raw materials has become unavailable. Processing changes likewise arise from the attempt to improve quality in terms of sensory, nutritional, microbiological stability factors, to reduce costs or to improve manufacturing productivity. Packaging changes arise from considerations of product stability or other quality factors, e.g., a certain amount of oxygen permeability may insure that a fresh beef product remains red in color for improved visual appeal to consumers. Packages function as carriers of product information and brand image, so both sensory characteristics and expectations can change as a function of how this information can be carried and displayed by the packaging material and its print overlay. Packaging and print ink may cause changes in flavor or aroma due to flavor transfer out of the product and sometimes transfer of off-flavors into the product. The package also serves as an important barrier to oxidative changes, to the potentially deleterious effects of light-catalyzed reactions, and to microbial infestations and other nuisances.

The sensory test is conducted to study how these product manipulations will create perceived changes to human observers. In this sense, sensory evaluation is in the best traditions of psychophysics, the oldest branch of scientific psychology, that attempts to specify the relationships between different energy levels impinging upon the sensory organs (the physical part of psychophysics) and the human response (the psychological part). Often, one cannot predict exactly what the sensory change will be as a function of ingredients, processes, or packaging, or it is very difficult to do so since foods and consumer products are usually quite complex systems. Flavors and aromas depend upon complex mixtures of many volatile chemicals. Informal tasting in the lab may not bring a reliable or sufficient answer to sensory questions. The benchtop in the development laboratory is a poor place to judge potential sensory impact with distractions, competing odors, nonstandard lighting, and so on. Finally, the nose, eyes, and tongue of the product developer may not be representative of most other people who will buy the product. So there is some uncertainty about how consumers will view a product especially under more natural conditions.

Uncertainty is the key here. If the outcome of a sensory test is perfectly known and predictable, there is no need to conduct the formal evaluation. Unfortunately, useless tests are often requested of a sensory testing group in the industrial setting. The burden of useless routine tests arises from overly entrenched product development sequences, corporate traditions, or merely the desire to protect oneself from blame in the case of unexpected failures. However, the sensory test is only as useful as the amount of reduction in uncertainty that occurs. If there is no uncertainty, there is no need for the sensory test. For example, doing a sensory test to see if there is a perceptible color difference between a commercial red wine and a commercial white wine is a waste of resources, since there is no uncertainty! In the industrial setting, sensory evaluation provides a conduit for information that is useful in management business decisions about directions for product development and product changes. These decisions are based on lower uncertainty and lower risk once the sensory information is provided.

Sensory evaluation also functions for other purposes. It may be quite useful or even necessary to include sensory analyses in quality control (QC) or quality assurance. Modification of traditional sensory practices may be required to accommodate the small panels and rapid assessments often required in on-line QC in the manufacturing environment. Due to the time needed to assemble a panel, prepare samples for testing, analyze and report sensory data, it can be quite challenging to apply sensory techniques to quality control as an on-line assessment. Quality assurance involving sensory assessments of finished products are more readily amenable to sensory testing and may be integrated with routine programs for shelf life assessment or quality monitoring. Often it is desirable to establish correlations between sensory response and instrumental measures. If this is done well, the instrumental measure can sometimes be substituted for the sensory test. This is especially applicable under conditions in which rapid turnaround is needed. Substitution of instrumental measurements for sensory data may also be useful if the evaluations are likely to be fatiguing to the senses, repetitive, involve risk in repeated evaluations (e.g., insecticide fragrances), and are not high in business risk if unexpected sensory problems arise that were missed.

In addition to these product-focused areas of testing, sensory research is valuable in a broader context. A sensory test may help to understand the attributes of a product that consumers view as critical to product acceptance and thus success. While we keep a wary eye on the fuzzy way that consumers use language, consumer sensory tests can provide diagnostic information about a product's points of superiority or shortcomings. Consumer sensory evaluations may suggest hypotheses for further inquiry such as exploration of new product opportunities.

There are recurrent themes and enduring problems in sensory science. In 1989, the ASTM Committee E-18 on Sensory Evaluation of Materials and Products published a retrospective celebration of the origins of sensory methods and the committee itself (ASTM, 1989). In that volume, Joe Kamen, an early sensory worker with the Quartermaster Food and Container Institute, outlined nine areas of sensory research which were active 45 years ago. In considering the status of sensory science in the first decade of the twenty-first century, we find that these areas are still fertile ground for research activity and echo the activities in many sensory labs at the current time. Kamen (1989) identified the following categories:

(1) Sensory methods research. This aimed at increasing reliability and efficiency, including research into procedural details (rinsing, etc.) and the use of different experimental designs. Meiselman (1993), a later sensory scientist at the U.S. Army Food Laboratories, raised a number of methodological issues then and even now still unsettled within the realm of sensory evaluation. Meiselman pointed to the lack of focused methodological research aimed at issues of measurement quality such as reliability, sensitivity, and validity. Many sensory techniques originate from needs for practical problem solving. The methods have matured to the status of standard practice on the basis of their industrial track record, rather than a connection to empirical data that compare different methods. The increased rate of experimental publications devoted to purely methodological comparisons in journals such as the Journal of Sensory Studies and Food Quality and Preference certainly points to improvement in the knowledge base about sensory testing, but much remains to be done.

(2) Problem solving and trouble shooting. Kamen raised the simple example of establishing product equivalence between lots, but there are many such day-to-day product-related issues that arise in industrial practice. Claim substantiation (ASTM E1958, 2008; Gacula, 1991) and legal and advertising challenges are one example. Another common example would be identification of the cause of off-flavors, "taints" or other undesirable sensory characteristics and the detective exercise that goes toward the isolation and identification of the causes of such problems.

(3) Establishing test specifications. This can be important to suppliers and vendors, and also for quality

control in multi-plant manufacturing situations, as well as international product development and the problem of multiple sensory testing sites and panels.

(4) Environmental and biochemical factors. Kamen recognized that preferences may change as a function of the situation (food often tastes better outdoors and when you are hungry). Meiselman (1993) questioned whether sufficient sensory research is being performed in realistic eating situations that may be more predictive of consumer reactions, and recently sensory scientists have started to explore this area of research (for example, Giboreau and Fleury, 2009; Hein et al., 2009; Mielby and Frøst, 2009).

(5) Resolving discrepancies between laboratory and field studies. In the search for reliable, detailed, and precise analytical methods in the sensory laboratory, some accuracy in predicting field test results may be lost. Management must be aware of the potential of false positive or negative results if a full testing sequence is not carried out, i.e., if shortcuts are made in the testing sequence prior to marketing a new product. Sensory evaluation specialists in industry do not always have time to study the level of correlation between laboratory and field tests, but a prudent sensory program would include periodic checks on this issue.

(6) Individual differences. Since Kamen's era, a growing literature has illuminated the fact that human panelists are not identical, interchangeable measuring instruments. Each comes with different physiological equipment, different frames of reference, different abilities to focus and maintain attention, and different motivational resources. As an example of differences in physiology, we have the growing literature on specific anosmias—smell "blindnesses" to specific chemical compounds among persons with otherwise normal senses of smell (Boyle et al., 2006; Plotto et al., 2006; Wysocki and Labows, 1984). It should not be surprising that some olfactory characteristics are difficult for even trained panelists to evaluate and to come to agreement (Bett and Johnson, 1996).

(7) Relating sensory differences to product variables. This is certainly the meat of sensory science in industrial practice. However, many product developers do not sufficiently involve their sensory specialists in the underlying research questions. They also may fall into the trap of never ending sequences of paired tests, with little or no planned designs and no modeling of how underlying physical variables (ingredients, processes) create a dynamic range of sensory changes. The relation of graded physical changes to sensory response is the essence of psychophysical thinking.

(8) Sensory interactions. Foods and consumer products are multidimensional. The more sensory scientists understand interactions among characteristics such as enhancement and masking effects, the better they can interpret the results of sensory tests and provide informed judgments and reasoned conclusions in addition to reporting just numbers and statistical significance.

(9) Sensory education. End-users of sensory data and people who request sensory tests often expect one tool to answer all questions. Kamen cited the simple dichotomy between analytical and hedonic testing (e.g., discrimination versus preference) and how explaining this difference was a constant task. Due to the lack of widespread training in sensory science, the task of sensory education is still with us today, and a sensory professional must be able to explain the rationale behind test methods and communicate the importance and logic of sensory technology to non-sensory scientists and managers.

1.3.1 Differences from Marketing Research Methods

Another challenge to the effective communication of sensory results concerns the resemblance of sensory data to those generated from other research methods. Problems can arise due to the apparent similarity of some sensory consumer tests to those conducted by marketing research services. However, some important differences exist as shown in Table 1.3. Sensory tests are almost always conducted on a blind-labeled basis. That is, product identity is usually obscured other than the minimal information that allows the product to be evaluated in the proper category (e.g., cold breakfast cereal). In contrast, marketing research tests often deliver explicit concepts about a product—label claims, advertising imagery, nutritional information, or any other information that may enter into the mix

Table 1.3 Contrast of sensory evaluation consumer tests with market research tests

Sensory testing with consumers
Participants screened to be users of the product category
Blind-labeled samples—random codes with minimal conceptual information
Determines if sensory properties and overall appeal met targets
Expectations based on similar products used in the category
Not intended to assess response/appeal of product concept
Market research testing (concept and/or product test)
Participants in product-testing phase selected for positive response to concept
Conceptual claims, information, and frame of reference are explicit
Expectations derived from concept/claims and similar product usage
Unable to measure sensory appeal in isolation from concept and expectations

designed to make the product conceptually appealing (e.g., bringing attention to convenience factors in preparation).

In a sensory test all these potentially biasing factors are stripped away in order to isolate the opinion based on sensory properties only. In the tradition of scientific inquiry, we need to isolate the variables of interest (ingredients, processing, packaging changes) and assess sensory properties as a function of these variables, and not as a function of conceptual influences. This is done to minimize the influence of a larger cognitive load of expectations generated from complex conceptual information. There are many potential response biases and task demands that are entailed in "selling" an idea as well as in selling a product. Participants often like to please the experimenter and give results consistent with what they think the person wants. There is a large literature on the effect of factors such as brand label on consumer response. Product information interacts in complex ways with consumer attitudes and expectancies (Aaron et al., 1994; Barrios and Costell, 2004; Cardello and Sawyer, 1992; Costell et al., 2009; Deliza and MacFie, 1996; Giménez et al., 2008; Kimura et al., 2008; Mielby and Frøst, 2009; Park and Lee, 2003; Shepherd et al., 1991/1992). Expectations can cause assimilation of sensory reactions toward what is expected under some conditions and under other conditions will show contrast effects, enhancing differences when expectations are not met (Siegrist and Cousin, 2009; Lee et al., 2006; Yeomans et al., 2008; Zellner et al., 2004). Packaging and brand information will also affect sensory judgments (Dantas et al., 2004; Deliza et al., 1999; Enneking et al., 2007). So the apparent resemblance of a blind sensory test and a fully concept-loaded market research test are quite illusory. Corporate management needs to be reminded of this important distinction. There continues to be tension between the roles of marketing research and sensory research within companies. The publication by Garber et al. (2003) and the rebuttal to that paper by Cardello (2003) are a relatively recent example of this tension.

Different information is provided by the two test types and both are very important. Sensory evaluation is conducted to inform product developers about whether they have met their sensory and performance targets in terms of perception of product characteristics. This information can only be obtained when the test method is as free as possible from the influences of conceptual positioning. The product developer has a right to know if the product meets its sensory goals just as the marketer needs to know if the product meets its consumer appeal target in the overall conceptual, positioning, and advertising mix. In the case of product failures, strategies for improvement are never clear without both types of information.

Sometimes the two styles of testing will give apparently conflicting results (Oliver, 1986). However, it is almost never the situation that one is "right" and the other is "wrong." They are simply different types of evaluations and are even conducted on different participants. For example, taste testing in market research tests may be conducted only on those persons who previously express a positive reaction to the proposed concept. This seems reasonable, as they are the likely purchasers, but bear in mind that their product evaluations are conducted *after they have already expressed some positive attitudes* and people like to be internally consistent. However, a blind sensory consumer test is conducted on a sample of regular product user, with no prescreening for conceptual interest or attitudes. So they are not necessarily the same sample population in each style of test and differing results should not surprise anyone.

1.3.2 Differences from Traditional Product Grading Systems

A second arena of apparent similarity to sensory evaluation is with the traditional product quality grading systems that use sensory criteria. The grading of agricultural commodities is a historically important influence on the movement to assure consumers of quality standards in the foods they purchase. Such techniques were widely applicable to simple products such as fluid milk and butter (Bodyfelt et al., 1988, 2008), where an ideal product could be largely agreed upon and the defects that could arise in poor handling and processing gave rise to well-known sensory effects. Further impetus came from the fact that competitions could be held to examine whether novice judges-in-training could match the opinions of experts. This is much in the tradition of livestock grading—a young person could judge a cow and receive awards at a state fair for learning to use the same criteria and critical eye as the expert judges. There are noteworthy differences in the ways in which sensory testing and quality judging are performed. Some of these are outlined in Table 1.4.

The commodity grading and the inspection tradition have severe limitations in the current era of highly processed foods and market segmentation. There are fewer and fewer "standard products" relative to the wide variation in flavors, nutrient levels (e.g., low fat), convenience preparations, and other choices that line the supermarket shelves. Also, one person's product defect may be another's marketing bonanza, as in the glue that did not work so well that gave us the ubiquitous post-it notes. Quality judging methods are poorly suited to research support programs. The techniques have been widely criticized on a number of scientific grounds (Claassen and Lawless, 1992; Drake, 2007; O'Mahony, 1979; Pangborn and Dunkley, 1964; Sidel et al., 1981), although they still have their proponents in industry and agriculture (Bodyfelt et al., 1988, 2008).

The defect identification in quality grading emphasizes root causes (e.g., oxidized flavor) whereas the descriptive approach uses more elemental singular terms to describe perceptions rather than to infer causes. In the case of oxidized flavors, the descriptive analysis panel might use a number of terms (oily, painty, and fishy) since oxidation causes a number of qualitatively different sensory effects. Another notable difference from mainstream sensory evaluation is that the quality judgments combine an overall quality scale (presumably reflecting consumer dislikes) with diagnostic information about defects, a kind of descriptive analysis looking only at the negative aspects of products. In mainstream sensory evaluation, the descriptive function and the consumer evaluation would be clearly separate in two distinct tests with different respondents. Whether the opinion of a single expert can effectively represent consumer opinion is highly questionable at this time in history.

Table 1.4 Contrast of sensory evaluation tests with quality inspection

Sensory testing
Separates hedonic (like–dislike) and descriptive information into separate tests
Uses representative consumers for assessment of product appeal (liking/disliking)
Uses trained panelists to specify attributes, but not liking/disliking
Oriented to research support
Flexible for new, engineered, and innovative products
Emphasizes statistical inference for decision making, suitable experimental designs, and sample sizes
Quality inspection
Used for pass–fail online decisions in manufacturing
Provides quality score and diagnostic information concerning *defects* in one test
Uses sensory expertise of highly trained individuals
May use only one or very few trained experts
Product knowledge, potential problems, and causes are stressed
Traditional scales are multi-dimensional and poorly suited to statistical analyses
Decision-making basis may be qualitative
Oriented to standard commodities