# Chapter 7

# Scaling

**Abstract** Scaling describes the application of numbers, or judgments that are converted to numerical values, to describe the perceived intensity of a sensory experience or the degree of liking or disliking for some experience or product. Scaling forms the basis for the sensory method of descriptive analysis. A variety of methods have been used for this purpose and with some caution, all work well in differentiating products. This chapter discusses theoretical issues as well as practical considerations in scaling.

*The vital importance of knowing the properties and limitations of a measuring instrument can hardly be denied by most natural scientists. However, the use of many different scales for sensory measurement is common within food science; but very few of these have ever been validated. . . .*
—(Land and Shepard, 1984, pp. 144–145)

## Contents



## 7.1 Introduction

People make changes in their behavior all the time based on sensory experience and very often this involves a judgment of how strong or weak something feels. We add more sugar to our coffee if it is not sweet enough. We adjust the thermostat in our home if it is too cold or too hot. If a closet is too dark to find your shoes you turn the light on. We apply more force to

H.T. Lawless, H. Heymann, *Sensory Evaluation of Food*, Food Science Text Series,
DOI 10.1007/978-1-4419-6488-5_7, © Springer Science+Business Media, LLC 2010

chew a tough piece of meat if it will not disintegrate to allow swallowing. These behavioral decisions seem automatic and do not require a numerical response. But the same kinds of experiences can be evaluated with a response that indicates the strength of the sensation. What was subjective and private becomes public data. The data are quantitative. This is the basis of scaling.

The methods of scaling involve the application of numbers to quantify sensory experiences. It is through this process of numerification that sensory evaluation becomes a quantitative science subject to statistical analysis, modeling, prediction, and hard theory. However, as noted in the quote above, in the practical application of sensory test methods, the nature of this process of assigning numbers to experiences is rarely questioned and deserves scrutiny. Clearly numbers can be assigned to sensations by a panelist in a variety of ways, some by mere categorization, or by ranking or in ways that attempt to reflect the intensity of sensory experience. This chapter will illustrate these techniques and discuss the arguments that have been raised to substantiate the use of different quantification procedures.

Scaling involves sensing a product or stimulus and then generating a response that reflects how the person perceives the intensity or strength of one or more of the sensations generated by that product. This process is based on a psychophysical model (see Chapter 2). The psychophysical model states that as the physical strength of the stimulus increases (e.g., the energy of a light or sound or the concentration of a chemical stimulus) the sensation will increase in some orderly way. Furthermore, panelists are capable of generating different responses to indicate these changes in what they experience. Thus a systematic relationship can be modeled of how physical changes in the real world result in changing sensations.

Scaling is a tool used for showing differences and degrees of difference among products. These differences are usually above the threshold level or just-noticeable difference. If the products are very similar and there is a question of whether there is any difference at all, the discrimination testing methods are more suitable (Chambers and Wolf, 1996). Scaling is usually done in one of the two scenarios. In the first, untrained observers are asked to give responses to reflect changes in intensity and it is presumed that (1) they understand the attribute they are asked to scale, e.g., sweetness and (2) there is no need to train or calibrate them to use the scale. This is the kind of scaling done to study a dose–response curve or psychophysical function. Such a study would perhaps be done on a student sample or a consumer population. A second kind of scaling is done when trained panelists are used as if they were measuring instruments, as in descriptive analysis. In this case they may be trained to insure a uniform understanding of the attribute involved (e.g., astringency) and often they are calibrated with reference standards to illustrate what is a little and what is a lot of this attribute. In this case the focus is in on the products being tested and not the more basic process of specifying a psychophysical function that has general application.

Note that these are "cheap data" (a term used by my advisor in graduate school, but perhaps "cost-effective" sounds a little less negative). One stimulus gives at least one data point. This is in contrast to indirect methods like forced choice tests. Many responses on a triangle test are needed to give one data point, i.e., the percent correct. Fechner and others referred to scaling as "the method of single stimuli" and considered it less reliable than the choice methods that were used to generate difference thresholds. However, direct scaling came into its own with the advent of magnitude estimation, an open-ended numerical response method. One or another type of scaling forms the basis for virtually all descriptive analysis techniques. In descriptive analysis, panelists generate scaled responses for various sensory attributes to reflect their subjective intensity.

There are two processes involved in scaling as shown in Fig. 7.1. The first is the psychophysical chain of events in which some energy or matter impinges upon receptors and the receptors send signals to the

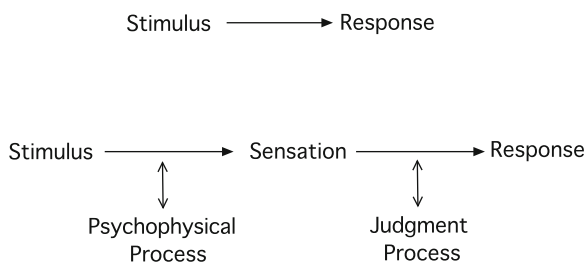


**Fig. 7.1** The two processes involved in scaling. The first, physiological process is the psychophysical translation of energy in the outside world into sensation, i.e., conscious experience. The second is the translation of that experience into some response. The psychophysical process can be modified by physiological processes such as adaptation and masking. The judgment function can be modified by cognitive processes such as contextual effects, number usage, and other response biases.

brain. These signals are interpreted in conscious perception as a sensation with some intensity or strength. The translation can be modified (i.e., the experience will change) by processes like adaptation or masking from another stimulus. The second process is the translation of that experience into an overt response (the data). This judgment function is influenced by the nature of the scaling task that the panelist is asked to perform. Factors such as contextual effects, the choice of comparison products, and response biases of that particular person can modify the process. The better the data reflect the experience, the more valid is the scaling method. The sensory professional must be careful to avoid response methods that introduce biases or non-sensory influences on the response output. For example, I might be asked to generate some open-ended numerical response to reflect my perception, but I might have some "favorite" numbers I find easy to use (e.g., integers or multiples of 2, 5, and 10) so this number bias interferes to some degree with the translation of my experience into a truly accurate response.

This chapter will focus on various methods that have been used in sensory evaluation and in psychophysics for scaling. Theory, principles, and issues will be discussed to provide depth of understanding. For the student who wishes to learn just the basic practices, Sections 7.3 and 7.4 are the most practically relevant sections. Section 7.5 illustrates some alternative methods that have appeared in the sensory evaluation literature, but have not at this time enjoyed widespread adoption in industrial practice. The sensory scientist should be aware of these additional methods for the potential advantages they may provide.

## 7.2 Some Theory

Measurement theory tells us that numbers can be assigned to items in different ways. This distinction was popularized by S. S. Stevens, the major proponent of magnitude estimation (1951). At least four ways of assigning numbers to events exist in common usage. These are referred to as nominal scaling, ordinal scaling, interval scaling, and ratio scaling.

In nominal scaling, numbers are assigned to events merely as labels. Thus gender may be coded as a "dummy variable" in statistical analysis by assigning a zero to males and a one to females; no assumption is made that these numbers reflect any ordered property of the sexes. They merely serve as convenient labels. The meals at which a food might be eaten could be coded with numbers as categories—one for breakfast, two for lunch, three for supper, and four for snacks. The assignment of a number for analysis is merely a label, a category or pigeonhole. The appropriate analysis of such data is to make frequency counts. The mode, the most frequent response, is used as a summary statistic for nominal data. Different frequencies of response for different products or circumstances can be compared by chi-square analysis or other nonparametric statistical methods (Siegel, 1956; see Appendix B). The only valid comparisons between individual items with this scale is to say whether they belong to the same category or to different ones (an equal versus not equal decision).

In ordinal scaling, numbers are assigned to recognize the rank order of products with regard to some sensory property, attitude, or opinion (such as preference). In this case increasing numbers assigned to the products represent increasing amounts or intensities of sensory experience. So a number of wines might be rank ordered for perceived sweetness or a number of fragrances rank ordered from most preferred to least preferred. In this case the numbers do not tell us anything about the relative differences among the products. We cannot draw conclusions about the degree of difference perceived nor the ratio or magnitude of difference. In an analogy to the order of runners finishing in a race, we know who placed first, second, third, etc. But this order does indicate neither the finishing distances between contestants nor the differences in their elapsed times. In general, analyses of ranked data can report medians as the summary statistic for central tendency or other percentiles to give added information. As with nominal data, nonparametric statistical analyses (see Appendix B) are appropriate when ranking is done (Siegel, 1956).

The next level of scaling occurs when the subjective spacing of responses is equal, so the numbers represent equal degrees of difference. This is called interval-level measurement. Examples in the physical sciences would be the centigrade and Fahrenheit scales of temperature. These scales have arbitrary zero points but equal divisions between values. The scales are inter-convertible through a linear transformation, for example, $°C = 5/9\ (°F–32)$. Few scales used in sensory science have been subjected to tests that would help establish whether they achieved an interval level of measurement and yet this level is often assumed.

One scale with approximately equal subjective spacing is the 9-point category scale used for like–dislike judgments, the 9-point hedonic scale (Peryam and Girardot, 1952). The phrases are shown below:

Like extremely
Like very much
Like moderately
Like slightly
Neither like nor dislike
Dislike slightly
Dislike moderately
Dislike very much
Dislike extremely

These response choices are commonly entered as data by assignment of the numbers one through nine. Extensive research was conducted to find the apparent spacing of various adjective labels for the scale points (Jones and Thurstone, 1955; Jones et al., 1955). The technique for deciding on the subjective spacing was the use of one kind of Thurstonian scaling method. The original data do not fully support the notion of equality of spacing, as discussed below. However, the scale worked well in practice, so this tradition of integer assignment has persisted. The method is illustrated in Appendix 1 at the end of this chapter. Thurstonian theory is discussed further in Chapter 5.

The advantage of interval-level measurement is that the data allow added interpretation. In a horse-racing example, we know the order the horses finished in and about how many "lengths" separated each horse. A second advantage is that more powerful statistical methods may be brought to bear the parametric methods. Computation of means, *t*-tests, linear regression, and analysis of variance are appropriate analyses.

Another even more desirable level of measurement is ratio measurement. In this case the zero level is fixed and not arbitrary and numbers will reflect relative proportions. This is the level of measurement commonly achieved in the physical sciences for quantities like mass, length, and temperature (on the absolute or Kelvin scale). Statements can be made that this item has twice as much length or mass than that item. Establishing whether a sensory scaling method actually assigns numbers to represent the relative proportions of different sensation intensities is a difficult matter. It has been widely assumed that the method of magnitude estimation is a priori a ratio scaling procedure. In magnitude estimation, subjects are instructed to assign numbers in relative proportions that reflect the strength of their sensations (Stevens, 1956). However, ratio instructions are easy to give, but whether the scale has ratio properties in reflecting a person's actual subjective experiences is difficult to determine, if not impossible.

Because of these different measurement types with different properties, the sensory professional must be careful about two things. First, statements about differences or ratios in comparing the scores for two products should not be made when the measurement level is only nominal or ordinal. Second, it is risky to use parametric statistics for measurements that reflect only frequency counts or rankings (Gaito, 1980; Townsend and Ashby, 1980). Nonparametric methods are available for statistical analyses of such data.

## 7.3 Common Methods of Scaling

Several different scaling methods have been used to apply numbers to sensory experience. Some, like magnitude estimation, are adapted from psychophysical research, and others, like category scaling have become popular through practical application and dissemination in a wide variety of situations. This section illustrates the common techniques of category scales, line marking, and magnitude estimation. The next section discusses the less frequently used techniques of hybrid category–ratio scales, indirect scales, and ranking as alternatives. Two other methods are illustrated. Intensity matching across sensory modalities, called cross-modality matching, was an important psychophysical technique and a precedent to some of the category–ratio scales. Finally, adjustable rating techniques in which panelists make relative placements and are able to alter their ratings are also discussed.

### *7.3.1 Category Scales*

Perhaps the oldest method of scaling involves the choice of discrete response alternatives to signify increasing sensation intensity or degrees of liking and/or preference. The alternatives may be presented in a horizontal or vertical line and may offer choices of integer numbers, simple check boxes, or word phrases. Examples of simple category scales are shown in

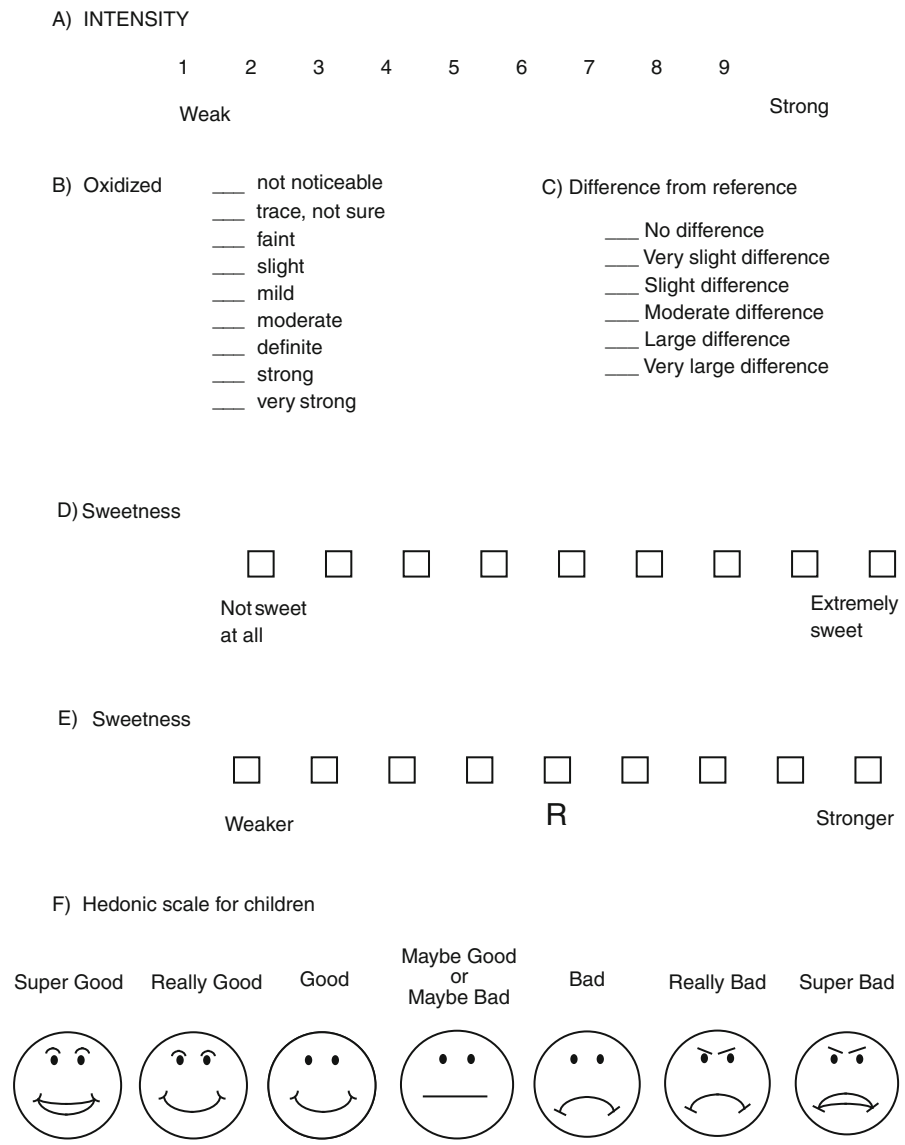


**Fig. 7.2** Examples of category scales. (**a**) a simple integer scale for sensation strength (after Lawless and Malone, 1986b); (**b**) a verbal scale for degree of oxidized flavor (after Mecredy et al., 1974; (**c**) a verbal scale for degree of difference from some reference or control sample (after Aust et al., 1985), (**d**) a simple check-box scale for perceived intensity; (**e**) a simple check-box scale for difference in intensity from some reference sample, marked R (after Stoer and Lawless, 1993); (**f**) a facial scale suitable for use with children, after Chen et al. (1996).

Fig. 7.2. The job of the consumer or panelist is to choose the alternative that best represents their reaction or sensation. In a category scale the number of alternative responses is limited. Seven to 15 categories are commonly used for intensity scaling depending upon the application and the number of gradations that the panelists are able to distinguish in the products. As panel training progresses, perceptual discrimination of intensity levels will often improve and more scale points may be added to allow the panel to make finer distinctions. A key idea is to present an easily understandable word like "sweetness" and ask the participant to evaluate the perceived intensity of that attribute. A second important factor concerns the verbal labels that appear along the alternatives. At the very least, the low and high ends of the scale must be labeled

with words that make sense, e.g., "not sweet at all" to "extremely sweet."

A wide variety of these scales have been used. A common version is to allow integer numerical responses of approximately nine points (e.g., Lawless and Malone, 1986a, b). Further gradations may be allowed. For example, Winakor et al. (1980) allowed options from 1 to 99 in rating attributes of fabric hand-feel. In the Spectrum method (Meilgaard et al., 2006) a 15-point category scale is used, but allows intermediate points in tenths, rendering it (at least in theory) a 150-point scale. In hedonic or affective testing, a bipolar scale is common, with a zero or neutral point of opinion at the center (Peryam and Girardot, 1952). These are often shorter than the intensity scales. For example, in the "smiling face" scale used with children, only three options may be used for very young respondents (Birch et al., 1980, 1982), although with older children as many as nine points may be used (Chen et al., 1996; Kroll, 1990). Lately there has been a move away from using labels or integers, in case these may be biasing to subjects. People seem to have favorite numbers or tendencies to use some numbers more than others (e.g., Giovanni and Pangborn, 1983). A solution to this problem is to use an unlabeled check-box scale as shown in Fig. 7.2.

In early applications of category scaling, the procedure specifically instructed subjects to use the categories to represent equal spacing. They might also be instructed to distribute their judgments over the available scale range, so the strongest stimulus was rated at the highest category and the weakest stimulus at the lowest category. This use of such explicit instructions surfaces from time to time. An example is in Anderson's (1974) recommendation to show the subject specific examples of bracketing stimuli that are above and below the anticipated range of items in the set to be judged. A related method is the relative scaling procedure of Gay and Mead (1992) in which subjects place the highest and lowest stimuli at the scale endpoints (discussed below). The fact that there is an upper boundary to the allowable numbers in a category scaling task may facilitate the achievement of a linear interval scale (Banks and Coleman, 1981).

A related issue concerns what kind of experience the high end anchor refers to. Muñoz and Civille (1998) pointed out that for descriptive analysis panels, the high end-anchor phrase could refer to different situations. For example, is the term "extremely sweet" referring to all products (a so-called universal scale)? Or is the scale anchored in the panelists' minds only to products in this category? In that case, extremely sweet for a salt cracker refers to something different than extremely sweet for a confectionary product or ice cream. Or is the high end of the scale the most extreme attribute for this product? That would yield a product-specific scale in which comparisons between different attributes could be made, e.g., this cracker is much sweeter than it is salty, but not comparisons to another type of product. These are important concerns for a descriptive panel leader.

However, some experimenters nowadays avoid any extra instructions, allowing the subject or panelist to distribute their ratings along the scale as they see fit. In fact, most people will have a tendency to distribute their judgments along most of the scale range, although some avoid the end categories, reserving them in case extreme examples show up. However, panelists do not like to overuse one part of the scale and will tend to move these judgments into adjoining response categories (Parducci, 1965). These tendencies are discussed in Chapter 9.

In practice, simple category scales are about as sensitive to product differences as other scaling techniques, including line marking and magnitude estimation (Lawless and Malone, 1986a, b). Due to their simplicity, they are well suited to consumer work. In addition, they offer some advantages in data coding and tabulation for speed and accuracy as they are easier to tabulate than measuring lines or recording the more variable magnitude estimates that may include fractions. This presumes, of course, that the data are being tabulated manually. If the data are recorded online using a computer-assisted data collection system, this advantage vanishes. A wide variety of scales with fixed alternatives are in current use. These include Likert-type scales used for opinions and attitudes, which are based on the degree to which a person agrees or disagrees with a statement about the product. Examples of such scales used in consumer tests are found in Chapter 14. Lately the term "Likert scale" has been used to refer to any category type of scale, but we prefer to reserve the name of Likert to the agree/disagree attitude scale in keeping with his original method (Likert, 1932). The flexibility of categorical alternatives for many different situations is thus one important aspect of the appeal of this kind of response measurement.

### 7.3.2 Line Scaling

A second widely used technique for intensity scaling involves making a mark or slash on a line to indicate the intensity of some attribute. Marking a line is also referred to as using a graphic-rating scale or a visual analog scale. The response is recorded as the distance of the mark from one end of the scale, usually whatever end is considered "lower." Line marking differs from category scales in the sense that the person's choices seem more continuous and less limited. In fact, the data are limited to the discrete choices measurable by the data-encoding instrument, such as the resolution of a digitizer or the number of pixels resolvable on a computer screen. The fundamental idea is that the panelist makes a mark on a line to indicate the intensity or amount of some sensory characteristic. Usually only the endpoints are labeled and marked with short line segments at right angles to the main line. The end-anchor lines may be indented to help avoid end effects associated with the reluctance of subjects to use the ends of scale. Other intermediate points may be labeled. One variation uses a central reference point representing the value of a standard or baseline product on the scale. Test products are scaled relative to that reference. Some of these variations are shown in Fig. 7.3. These techniques are very popular in descriptive analysis in which multiple attributes are evaluated by trained panels.

The first use of line scales for sensory evaluation appears in an experiment from the Michigan State Agricultural Experiment Station conducted during World War II (Baten, 1946). Various storage temperatures of apples were tested. A simple category scale for fruit appeal was used (ratings from very desirable to very undesirable with seven alternatives) and then a 6 in. line scale was used, with the words "very poor" over the left end and "excellent" over the right end. Responses on the line were measured in inches. A poll of the participants revealed a strong preference for the category scale over the line marking scale. However, Baten reported that the *t*-values comparing apples were about twice as large using the line-marking technique, implying a greater ability to statistically differentiate the items using the line scale. Unfortunately, Baten did not report any numerical values for the *t*-statistics, so it is difficult to evaluate the size of the advantage he saw.

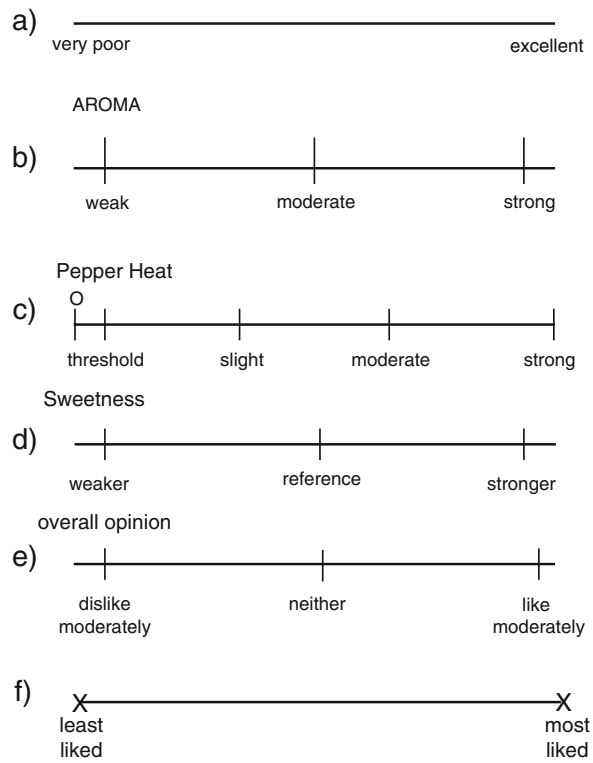


**Fig. 7.3** Examples of line-marking scales: (**a**) with endpoints labeled (after Baten, 1946); (**b**) with indented "goal posts" (after Mecredy et al., 1974); (**c**) with additional points labeled as in ASTM procedure E-1083 (ASTM, 2008b); (**d**) a line for ratings relative to a reference as in Stoer and Lawless (1993); (**e**) hedonic scaling using a line; (**f**) the adjustable scale of Gay and Mead (1992) as pictured by Villanueva and D Silva (2009).

An important historical trend to use lines in descriptive analysis was instrumental in the popularization of line marking. Stone et al. (1974) recommended the use of line marking for Quantitative Descriptive Analysis (QDA), then a relatively new approach to specifying the intensities of all important sensory attributes. It was important to have a scaling method in QDA that approximated an interval scale, as analysis of variance was to become the standard statistical technique for comparing products in descriptive analysis. The justification for the application in QDA appears to rest on the previous findings of Baten regarding the sensitivity of the method and the writings of Norman Anderson on functional measurement theory (Anderson, 1974). In his approach, Anderson used indented end anchors (see Weiss, 1972, for another example). Anderson also showed his subjects examples of the high and

low stimuli that one might encounter in the experiment, and sometimes even more extreme examples, in order to orient and stabilize their scale usage. Whether such examples make sense for sensory evaluation is questionable.

Since the advent of QDA, the line-marking techniques have been applied in many different situations requiring sensory response. In an early sensory application, Einstein (1976) successfully used a line-marking scale with consumers to evaluate beer attributes of flavor intensity, body, bitterness, and aftertaste. By "successful" we mean that statistically significant differences were obtained among test samples. The use of line marking is not limited to foods and consumer products. Measurement of pain and pain relief has employed line marking in clinical settings, using both vertical and horizontal lines (Huskisson, 1983; Sriwatanakul et al., 1983). Lawless (1977) used a line-marking technique along with ratio instructions for both intensity and hedonic judgments in taste and odor mixture studies. This was a hybrid procedure in which subjects were instructed to mark lines as if performing magnitude estimation. For example, if one product were twice as sweet as a previous item, a mark would be made twice as far down the line (which could be extended if the panelist ran out of room). Villanueva and colleagues used a scale with equally spaced dots along the line and obtained good results for acceptability scaling (Villanueva and Da Silva, 2009; Villanueva et al., 2005). In comparisons of category ratings, line marking, and magnitude estimation, the line-marking method is about as sensitive to product differences as other scaling techniques (Lawless and Malone, 1986a, b).

Marking a point on a line has also been used widely in time–intensity scaling methods. The simplest version of this is to move a pointer along a scale while a moving roll of paper is marked to see the continuous changes in sensation over time. Originally this could be done with a simple marking pen held by the participant (e.g., Moore and Shoemaker, 1981). The record of the pen track would usually be obscured from the person's view, so as not to exert any influence on their response. The pen-tracking method has also been used with ratio instructions (Lawless and Skinner, 1979). In some cases, the participant has turned a dial or other response device while observing a linear display (Lawless and Clark, 1992). Often the time–intensity scale will look much like a vertical "thermometer" with a cursor that moves up and down via the computer mouse. Time–intensity methods are reviewed more fully in Chapter 8.

## 7.3.3 Magnitude Estimation

### 7.3.3.1 The Basic Techniques

A popular technique for scaling in psychophysical studies has been the method of magnitude estimation. In this procedure, the respondent is instructed to assign numbers to sensations in proportion to how strong the sensation seems. Specifically, the ratios between the numbers are supposed to reflect the ratios of sensation magnitudes that have been experienced. For example, if product A is given the value of 20 for sweetness intensity and product B seems twice as sweet, B is given a magnitude estimate of 40. The two critical parts of the technique are the instructions given to the participant and the techniques for data analysis. Two primary variations of magnitude estimation have been used. In one method, a standard stimulus is given to the subject as a reference and that standard is assigned a fixed value such as 10. All subsequent stimuli are rated relative to this standard, sometimes called a "modulus." It is often easier for panelists if the reference (i.e., the item used as the modulus) is chosen from somewhere near the middle of the intensity range.

In the other variation of magnitude estimation, no standard stimulus is given and the participant is free to choose any number he or she wishes for the first sample. All samples are then rated relative to this first intensity, although in practice people probably "chain" their ratings to the most recent items in the series. Because people can choose different ranges of numbers in this "non-modulus" magnitude estimation, the data have to be treated to bring all judgments into the same range, an extra step in the analysis. Variations on magnitude estimation and guidelines for data analysis are found in ASTM Standard Test Method E 1697–05 (ASTM, 2008a).

In the psychophysical laboratory, where magnitude estimation has found its primary usage, generally only one attribute is rated at a time. However, rating multiple attributes or profiling has been used in taste studies (McBurney and Bartoshuk, 1973; McBurney and Shick, 1971; McBurney et al., 1972) and this can

naturally be extended to foods with multiple taste and aromatic attributes. Magnitude estimation has not been used very often for descriptive analysis, but in principle there is no reason why it could not be used for that purpose.

Participants should be cautioned to avoid falling into previous habits of using bounded category scales, e.g., limited ranges of numbers from zero to ten. This may be a difficult problem with previously trained panels that have used a different scaling method, as people like to stick with a method they know and feel comfortable with. Panelists who show such behavior may not understand the ratio nature of the instructions. It is sometimes useful with a new panelist to have the participant perform a warm-up task to make sure they understand the scaling instructions. The warm-up task can involve estimation of the size or area of different geometric figures (Meilgaard et al., 2006) or the length of lines (McBurney et al., 1972). Sometimes it is desired to have panelists rate multiple attributes at the same time or to break down overall intensity into specific qualities. If this kind of "profiling" is needed, the geometric figures can include areas with different shading or the lines can be differently colored. A practice task is highly recommended so that the sensory scientist can check on whether the participant understands the task.

Values of zero are allowed in this method as some of the products may in fact have or no sensation for a given attribute (like no sweetness in our example). Of course, the rating of zero should not be used for the reference material. While the value of zero is consistent with common sense for products with no sensation of some attributes, it does complicate the data analysis as discussed below.

#### 7.3.3.2 Instructions

The visual appearance of the ballot in magnitude estimation is not critical; it is the instructions and the participant's comprehension of the ratio nature of the judgments that are important. Some ballots even allow the subject/participant to view all previous ratings. Here are sample instructions for the use of magnitude estimation with a reference sample or modulus with a fixed number assigned to it:

> Please taste the first sample and note its sweetness. This sample is given the value of "10" for its sweetness intensity. Please rate all other samples in proportion to this reference. For example, if the next sample is twice as sweet, assign it a value of "20", if half as sweet, assign it a value of "5" and if 3.5 times as sweet, assign in a value of 35. In other words, rate the sweetness intensity so that your numbers represent the ratios among the intensities of sweetness. You may use any positive numbers including fractions and decimals.

The other major variation on this method uses no reference. In this case the instructions may read as follows:

> Please taste the first sample and note its sweetness. Please rate all other samples relative to this reference, applying numbers to the samples to represent the ratios of sweetness intensity among the samples. For example, if the next sample was twice as sweet, you would give it a number twice as big as the rating assigned to the first sample, if half as sweet, assign it a number half as big and if 3.5 times as sweet, assign it a number 3.5 times as big. You may use any positive numbers including fractions and decimals.

#### 7.3.3.3 Data Treatment

In non-modulus methods, participants will generally choose some range of numbers they feel comfortable with. The ASTM procedure suggests having them pick a value between 30 and 100 for the first sample, and avoiding any number that seems small. If participants are allowed to choose their own number range, it becomes necessary to re-scale each individual's data to bring them into a common range before statistical analysis (Lane et al., 1961). This will prevent subjects who choose very large numbers from having undue influence on measures of central tendency (means) and in statistical tests. This rescaling process has been referred to as "normalizing" (ASTM, 2008a) although it has nothing to do with the normal distribution or *Z*-scores. A common method for rescaling proceeds as follows:

(1) Calculate the geometric mean of each individual's ratings across their data set.
(2) Calculate the geometric mean of the entire data set (of all subjects combined).
(3) For each subject, construct a ratio of the grand geometric mean of the entire data set to each person's geometric mean. The value of this ratio provides a post hoc individual rescaling factor for each subject. In place of the grand geometric mean,

any positive numerator may also be chosen in constructing this factor, e.g., a value of 100.

(4) Multiply each data point for a given person by their individual rescaling factor. Do this for all participants using their own individual re-scaling factors.

These re-scaled data are then analyzed. Note that due to the extra data treatment step in this method, it is simpler to use the modulus-based variation with a standard reference item.

Magnitude estimation data are often transformed to logs before data analysis (Butler et al., 1987; Lawless, 1989). This is done primarily because the data tend to be positively skewed or log-normally distributed. There tends to be some high outlying values for any given sample. Perhaps this is not surprising because the scale is open-ended at the top, and bounded by zero at the bottom. Transformation into log data and/or taking geometric means presents some problems, however, when the data contain zeros. The log of zero is undefined. Any attempt to take a geometric mean by calculating the product of *N* items will yield a zero on multiplying. Several approaches have been taken to this problem. One is to assign a small positive value to any zeros in the data, perhaps one-half of the smallest rating given by a subject (ASTM, 2008a). The resulting analysis, however, will be influenced by this choice. Another approach is to use the median judgments in constructing the normalization factor for non-modulus methods. The median is less influenced by the high outliers in the data than the arithmetic mean.

#### 7.3.3.4 Applications

For practical purposes, the method of magnitude estimation may be used with trained panels, consumers, and even children (Collins and Gescheider, 1989). However, the data do tend to be a bit more variable than other bounded scaling methods, especially in the hands of untrained consumers (Lawless and Malone, 1986b). The unbounded nature of the scale may make it especially well suited to sensory attributes where an upper boundary might impose restrictions on the panelists' ability to differentiate very intense sensory experiences in their ratings. For example, irritative or painful sensations such as chili pepper intensity might all be rated near the upper bound of a category scale of intensity, but an open-ended magnitude estimation procedure would allow panelists more freedom to differentiate and report variations among very intense sensations.

With hedonic scaling of likes and dislikes, there is an additional decision in using magnitude estimation scaling. Two options have been adopted in using this technique, one employing a single continuum or unipolar scale for amount of liking and the other applying a bipolar scale with positive and negative numbers plus a neutral point (Pearce et al., 1986). In bipolar magnitude scaling of likes and dislikes, positive and negative numbers are allowed in order to signify ratios or proportions of both liking and disliking (e.g., Vickers, 1983). An alternative to positives and negatives is to have the respondent merely indicate whether the number represents liking or disliking (Pearce et al., 1986). In unipolar magnitude estimation only positive numbers (and sometimes zeros) are allowed, with the lower end of the scale representing no liking and higher numbers given to represent increasing proportions of liking (Giovanni and Pangborn, 1983; Moskowitz and Sidel, 1971). It is questionable whether a unipolar scale is a sensible response task for the participant, as it does not recognize the fact that a neutral hedonic response may occur, and that there are clearly two modes of reaction, one for liking and one for disliking. If one assumes that all items are on one side of the hedonic continuum—either all liked to varying degrees or all disliked to varying degrees then the one-sided scale makes sense. However, it is a rare situation with foods or consumer product testing in which at least some indifference or change of opinion was not visible in at least some respondents. So a bipolar scale fits common sense.

## 7.4 Recommended Practice and Practical Guidelines

Both line scales and category scales may be used effectively in sensory testing and consumer work. So we will not expend much effort in recommending one of these two common techniques over another. Some practical concerns are given below to help the student or practitioner avoid some potential problems. The category–ratio or labeled magnitude scales may facilitate comparisons among different groups, and this issue is discussed below in Section 7.5.2.

### 7.4.1 Rule 1: Provide Sufficient Alternatives

One major concern is to provide sufficient alternatives to represent the distinctions that are possible by panelists (Cox, 1980). In other words, a simple 3-point scale may not suffice if the panel is highly trained and capable of distinguishing among many levels of the stimuli. This is illustrated in the case of the flavor profile scale, which began with five points to represent no sensation, threshold sensation, weak, moderate, and strong (Caul, 1957). It was soon discovered that additional intermediate points were desirable for panelists, especially in the middle range of the scale where many products would be found. However, there is a law of diminishing returns in allowing too many scale points—further elaboration allows better differentiation of products up to a point and then the gains diminish as the additional response choices merely capture random error variation (Bendig and Hughes, 1953).

A related concern is the tendency, especially in consumer work, to simplify the scale by eliminating options or truncating endpoints. This brings in the danger caused by end-use avoidance. Some respondents may be reluctant to use the end categories, just in case a stronger or weaker item may be presented later in the test. So there is some natural human tendency to avoid the end categories. Truncating a 9-point scale to a 7-point scale may leave the evaluator with what is functionally only a 5-point scale for all practical purposes. So it is best to avoid this tendency to truncate scales in experimental planning.

### 7.4.2 Rule 2: The Attribute Must Be Understood

Intensity ratings must be collected on an attribute which the participants understand and about which they have a general consensus and agreement as to its meaning. Terms like sweetness are almost universal but a term like "green aroma" might be interpreted in different ways. In the case of a descriptive panel, a good deal of effort may be directed at using reference standards to illustrate what is meant by a specific term. In the case of consumer work, such training is not done, so if any intensity ratings are collected, they must be about simple terms about which people generally agree. Bear in mind that most early psychophysics was done on simple attributes like the loudness of a sound or the heaviness of a weight. In the chemical senses, with their diverse types of sensory qualities and fuzzy consumer vocabulary, this is not so straightforward.

Other problems to avoid include mixing sensation intensity (strength) and hedonics (liking), except in the just-right scale where this is explicit. An example of a hedonically loaded sensory term is the adjective "fresh." Whatever this means to consumers, it is a poor choice for a descriptive scale because it is both vague and connotes some degree of liking or goodness to most people. Vague terms are simply not actionable when it comes to giving feedback to product developers about what needs to be fixed. Another such vague term that is popular in consumer studies is "natural." Even though consumers might be able to score products on some unknown basis using this word, the information is not useful as it does not tell formulators what to change if a product scores low. A similar problem arises with attempting to scale "overall quality." Unless quality has been very carefully defined, it cannot be scaled.

### 7.4.3 Rule 3: The Anchor Words Should Make Sense

In setting up the scales for descriptive analysis or for a consumer test, the panel leader should carefully consider the nature of the verbal end anchors for each scale as well as any intermediate anchors that may be needed. Should the scale be anchored from "very weak" to "very strong" or will there be cases in which the sensory attribute is simply not present? If so, it makes sense to verbally anchor the bottom of the scale with "not at all" or "none." For example, a sweetness scale could be anchored with "not at all sweet" and a scale for "degree of oral irritation" could be anchored using "none."

### 7.4.4 To Calibrate or Not to Calibrate

If a high degree of calibration among the panelists is desired, then physical standards can be given for intensity. Often this is done with end examples as discussed above, but it may be advantageous to give

examples of intermediate points on the scales as well. An example of this kind of calibration is found in the ASTM procedures for evaluating pepper heat, where three points on the scale are illustrated (weak, moderate, and approaching strong) (ASTM, 2008b). The traditional texture profile technique (Brandt et al., 1963) used nine scale points for most texture attributes like hardness and would give examples of common products representing each point on the scale. In the Spectrum descriptive method, the scales for intensity are intended to be comparable across all attributes and all products so scale examples are given from various sensory domains representing points on the 15-point scale for intensity (Meilgaard et al., 2006). Whether or not this degree of calibration is required for a specific project should be considered. There may also be a limit to the ability to stabilize the scale usage of respondents. There are limits on the abilities of humans to be calibrated as measuring instruments (Olabi and Lawless, 2008) and in spite of decades of research in scaling, this is not well understood. People differ in their sensitivities to various tastes and odors and thus may honestly differ in their sensory responses.

Another decision of the test designer will be whether to assign physical examples to intermediate scale points. Although reference items are commonly shown for the end categories, this is less often done with intermediate categories. The apparent advantage is to achieve a higher level of calibration, a desirable feature for trained descriptive panelists. A potential disadvantage is the restriction of the subject's use of the scale. What appears to be equal spacing to the experimenter may not appear so to the participant. In that case, it would seem wiser to allow the respondent to distribute his or her judgments among the scale alternatives, without presuming that the examples of intermediate scale points are in fact equally spaced. This choice is up to the experimenter. The decision reflects one's concerns as to whether it is more desirable to work toward calibration or whether there is more concern with potential biasing or restriction of responses.

### 7.4.5 A Warning: Grading and Scoring are Not Scaling

In some cases pseudo-numerical scales have been set up to resemble category scales, but the examples cut across different sensory experiences, mixing qualities. An example is the pseudo-scale used for baked products, where the number 10 is assigned for perfect texture, 8 for slight dryness, 6 for gumminess, and 4 if very dry (AACC, 1986). Gumminess and dryness are two separate attributes and should be scaled as such. This is also an example of quality grading, which is not a true scaling procedure. When the numbers shift among different sensory qualities, this violates the psychophysical model for scaling the intensity of a single attribute. Although numbers may be applied to the grades, they cannot be treated statistically, as the average of "very dry " (4) and slightly dry (8) is not gummy (6) (see Pangborn and Dunkley, 1964, for a critique of this in the dairy grading arena). The numbers in a quality-grading scheme do not represent any kind of unitary psychophysical continuum.

## 7.5 Variations—Other Scaling Techniques

An important idea in scaling theory is the notion that people may have a general idea of how weak or strong sensations are, and that they can compare different attributes of a product for their relative strength, even across different sensory modalities. So, for example, someone could legitimately say that this product tastes much more salty than it is sweet. Or that the trumpets are twice as loud as the flutes in a certain passage in a symphony. Given that this notion is correct, people would seem to have a general internal scale for the strength of sensations. This idea forms the basis for several scaling methods. It permits the comparison of different sensations cross-referenced by their numerical ratings, and even can be used to compare word responses. Methods derived from this idea are discussed next.

### 7.5.1 Cross-Modal Matches and Variations on Magnitude Estimation

The method of magnitude estimation has a basis in earlier work such as fractionation methods and the method of sense ratios in the older literature (Boring,

1942) where people would be asked to set one stimulus in a given sensation ratio to another. The notion of allowing any numbers to be generated in response to stimuli, rather than adjusting stimuli to represent fixed numbers appeared somewhat later (Moskowitz, 1971; Richardson and Ross, 1930; Stevens, 1956). An important outcome of these studies was the finding that the resulting psychophysical function generally conformed to a power law of the following form:

$$R = kI^n \tag{7.1}$$

or after log transformation:

$$\log(R) = n\log(I) + \log(k) \tag{7.2}$$

where $R$ was the response, e.g., perceived loudness (mean or geometric mean of the data) and $I$ was the physical stimulus intensity, e.g., sound pressure, and $k$ was a constant of proportionality that depends upon the units of measurement. The important characteristic value of any sensory system was the value for $n$, the exponent of the power function or slope of the straight line in a log–log plot (Stevens, 1957). The validity of magnitude estimation then came to hang on the validity of the power law—the methods and resulting functions formed an internally consistent theoretical system. Stevens also viewed the method as providing a direct window into sensation magnitude and did not question the idea that these numbers generated by subjects might be biased in some way. However, the generation of responses is properly viewed as combining at least two processes, the psychophysical transformation of energy into conscious sensation and the application of numbers to those sensations. This response process was not given due consideration in the early magnitude estimation work. Responses as numbers can exhibit nonlinear transformations of sensation (Banks and Coleman, 1981; Curtis et al., 1968) so the notion of a direct translation from sensation to ratings is certainly a dangerous oversimplification.

Ratio-type instructions have been applied to other techniques as well as to magnitude estimation. A historically important psychophysical technique was that of cross-modality matching, in which the sensation levels or ratios would be matched in two sensory continua such as loudness and brightness. One continuum would be adjusted by the experimenter and the other by the subject. For example, one would try to make the brightness of the lights about in the same proportions as the loudness of the sounds. Stevens (1969) proposed that these experiments could validate the power law, since the exponents of the cross-modality matching function could be predicted from the exponents derived from separate scaling experiments. Consider the following example:

For one sensory attribute (using the log transform, Eq. (7.2)),

$$\log R_1 = n_1 \log I_1 + \log k_1 \tag{7.3}$$

and for a second sensory attribute

$$\log R_2 = n_2 \log I_2 + \log k_2 \tag{7.4}$$

Setting $R_1 = R_2$ in cross-modality matching gives

$$n_1 \log I_1 + \log k_1 = n_2 \log I_2 + \log k_2 \tag{7.5}$$

and rearranging,

$$\log I_1 = (n_2/n_1) \log I_2 + \text{a constant} \tag{7.6}$$

If one plots $\log I_1$ against $\log I_2$ from a cross-modality matching task, the slope of the function can be predicted from the ratio of the slopes of the individual exponents (i.e., $n_2/n_1$, which you can derive from two separate magnitude estimation tasks). This prediction holds rather well for a large number of compared sensory continua (Steven, 1969). However, whether it actually provides a validation for the power law or for magnitude estimation has been questioned (e.g., Ekman, 1964).

For practical purposes, it is instructive that people can actually take disparate sensory continua and compare them using some generalized notion of sensory intensity. This is one of the underpinnings of the use of a universal scale in the Spectrum descriptive procedure (Meilgaard et al., 2006). In that method, different attributes are rated on 15-point scales that can (in theory) *be meaningfully compared*. In other words, a 12 in sweetness is twice as intense a sensation as a 6 in saltiness. Such comparisons seem to makes sense for tastes and flavors but may not cut across all other modalities. For example, it might seem less sensible to compare the rating given for the amount of chocolate chips in a cookie to the rating given for the cookie's hardness—these seem like quite different experiences to quantify.

Cross-modality matches have been performed successfully with children even down to age 4, using line length compared to loudness (Teghtsoonian, 1980). This may have some advantage for those with limited verbal skills or trouble understanding the numerical concepts needed for category scales or magnitude estimation. The use of line length points out that the line-marking technique might be considered one form of a cross-modality matching scale. Some continua may seem simpler, easier, or more "natural" to be matched (for example, hand grip force to perceived strength of tooth pain). King (1986) matched the pitches of tones to concentrations of benzaldehyde and Ward (1986) used duration as a matching continuum for loudness and brightness. One of the advantages of the cross-modality matching procedure is that it is possible to specify the intensity of a sensation in physical units, i.e., as a physical level on the other continuum. So sweetness, for example, could be represented in a decibel (sound pressure) equivalent. In one amusing variation on this idea, Lindvall and Svensson (1974) used hedonic matching to specify the unpleasantness of combustion toilet fumes to different levels of $H_2S$ gas that were sniffed from an olfactometer. Thus the lucky participant could dial up a concentration that was perceived as being equally as offensive as the test samples.

If line marking can be considered a kind of cross-modality match, then why not the numbers themselves? It should be possible to cross-reference one continuum to another simply through instructions to use a common response scale. Stevens and Marks (1980) developed the technique of magnitude matching to do just that (see also Marks et al., 1992). Subjects were instructed to judge loudness and brightness on a common scale of intensity so that if a sound had the same sensory impact as the brightness of a light, then the stimuli would be given the same number. This type of cross-referencing should facilitate comparisons among people. For example, if it can be assumed that two people have the same response to salt taste or to loudness of tones, then differences in some other continuum like bitter taste, hot chili pepper intensity, or a potential olfactory loss can be cross-referenced through salty taste or through loudness of tones they have rated in magnitude matching (e.g., Gent and Bartoshuk, 1983). Furthermore, if numbers can provide a cross-referencing continuum, then why not scale the word phrases used as anchor points on a category or line scale? The idea of scaling word phrases takes shape in the labeled magnitude scales discussed next.

### 7.5.2 Category–Ratio (Labeled Magnitude) Scales

A group of hybrid techniques for scaling has recently enjoyed some popularity in the study of taste and smell, for hedonic measurement and other applications. One of the problems with magnitude estimation data is that it does not tell in any absolute sense whether sensations are weak or strong, only giving the ratios among them. This group of scales attempts to provide ratio information, but combines it with common verbal descriptors along a line scale to provide a simple frame of reference. They are referred to as category–ratio scales, or more recently, labeled magnitude scales. They all involve a horizontal or vertical line with deliberately spaced labels and the panelists' task is to make a mark somewhere along the line to indicate the strength of their perception or strength of their likes or dislikes. In general, these labeled line scales give data that are consistent with those from magnitude estimation (Green et al., 1993). An unusual characteristic of these scales is the verbal high end-anchor phrase, which often refers to the "strongest imaginable."

The technique is based on early work by Borg and colleagues, primarily in the realm of perceived physical exertion (Borg, 1982, 1990; see Green et al., 1993). In developing this scale, Borg assumed that the semantic descriptors could be placed on a ratio scale and that they defined the level of perceptual intensity and that all individuals experienced the same perceptual range. Borg suggested that for perceived exertion, the maximal sensation is roughly equivalent across people for this sensory "modality" (Marks et al., 1983). For example, it is conceivable that riding a bicycle to the point of physical exhaustion produces a similar sensory experience for most people. So the scale came to have the highest label referring to the strongest sensation imaginable.

This led to the development of the labeled magnitude scale (LMS) shown in Fig. 7.4. It is truly a hybrid method since the response is a vertical line-marking task but verbal anchors are spaced according to calibration using ratio-scaling instructions (Green

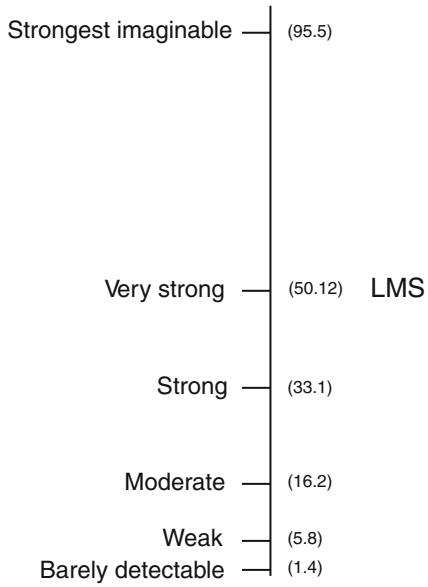


**Fig. 7.4** The labeled magnitude scale (LMS) of Green et al. (1993).

et al., 1993). In setting up the scale, Green and colleagues had subjects provide magnitude estimates of different verbal descriptors after giving magnitude estimates of familiar oral sensations (e.g., the bitterness of celery, the burn of cinnamon gum). These results were generally in line with previous scaling of verbal descriptors, so-called semantic scaling. Other researchers have developed scales with only direct scaling of the verbal descriptors and have not always included a list of everyday or common experiences (Cardello et al., 2003; Gracely et al., 1978a, b).

After the introduction of the LMS, a number of researchers tried to extend this approach into the realm of hedonics (measuring food acceptability). A widely used scale is the labeled affective magnitude (LAM) scale developed by Schutz and Cardello (2001). They used direct ratio scaling of the verbal descriptors of the 9-point hedonic scale and included Borg's type of high (and low) end anchor ("greatest imaginable like/dislike"). The scale is shown in Fig. 7.5. This shows some advantages in differentiating well-liked items (El Dine and Olabi, 2009; Greene et al., 2006; Schutz and Cardello, 2001), although that finding is not universal (Lawless et al., 2010a). The LAM scale or similar versions of it have been applied in a variety of studies with different foods (Chung and Vickers, 2007a, b; El Dine and Olabi, 2009; Forde and Delahunty, 2004; Hein et al., 2008; Keskitalo et al., 2007; Lawless et al., 2010a, b, c). A growing number of similar scales have been developed for various applications including oral pleasantness/unpleasantness (the "OPUS" scale, Guest et al., 2007), perceived satiety (the "SLIM" scale, Cardello et al., 2005), clothing fabric comfort (the "CALM" scale, Cardello et al., 2003), and odor dissimilarity (Kurtz et al., 2000). All of these scales depend upon a ratio scaling task to determine the spacing of the verbal descriptors and almost all use a Borg-type high end-anchor phrase. Others will surely be developed.

Instructions to participants have differed in the use of these scales. In the first application of the LMS, Green et al. (1993) instructed subjects to first choose the most appropriate verbal descriptor, and then to "fine tune" their judgment by placing a mark on the line between that descriptor and the next most appropriate one. In current practice less emphasis may be placed on the consideration of the verbal labels and instructions may be given to simply make a mark "anywhere" on the line. A common observation with the hedonic versions of the scale is that some panelists will mark at or very near a verbal descriptor, seeming to use it as a category scale (Cardello et al., 2008; Lawless et al., 2010a). The proportion of people displaying this behavior may depend upon the physical length of the line (and not the instructions or examples that may be shown) (Lawless et al., 2010b).

Results may depend in part on the nature of the high end-anchor example and the frame of reference of the subject in terms of the sensory modality they are thinking about. Green et al. (1996) studied the application of the LMS to taste and odor, using descriptors for the upper bound as "strongest imaginable" taste, smell, sweetness, etc. Steeper functions (a smaller response range) were obtained when mentioning individual taste qualities. This appears to be due to the omission of painful experiences (e.g., the "burn of hot peppers") from the frame of reference when sensations were scaled relative to only taste. The steepening of the functions for the truncated frame of reference is consistent with the idea that subjects expanded their range of numbers as seen in other scaling experiments (e.g., Lawless and Malone, 1986b). The fact that subjects appear to adjust their perceptual range depending on instructions or frame of reference suggests that the scales have relative and not absolute properties, like most other scaling methods. Cardello et al. (2008) showed that the hedonic version of the scale (the LAM

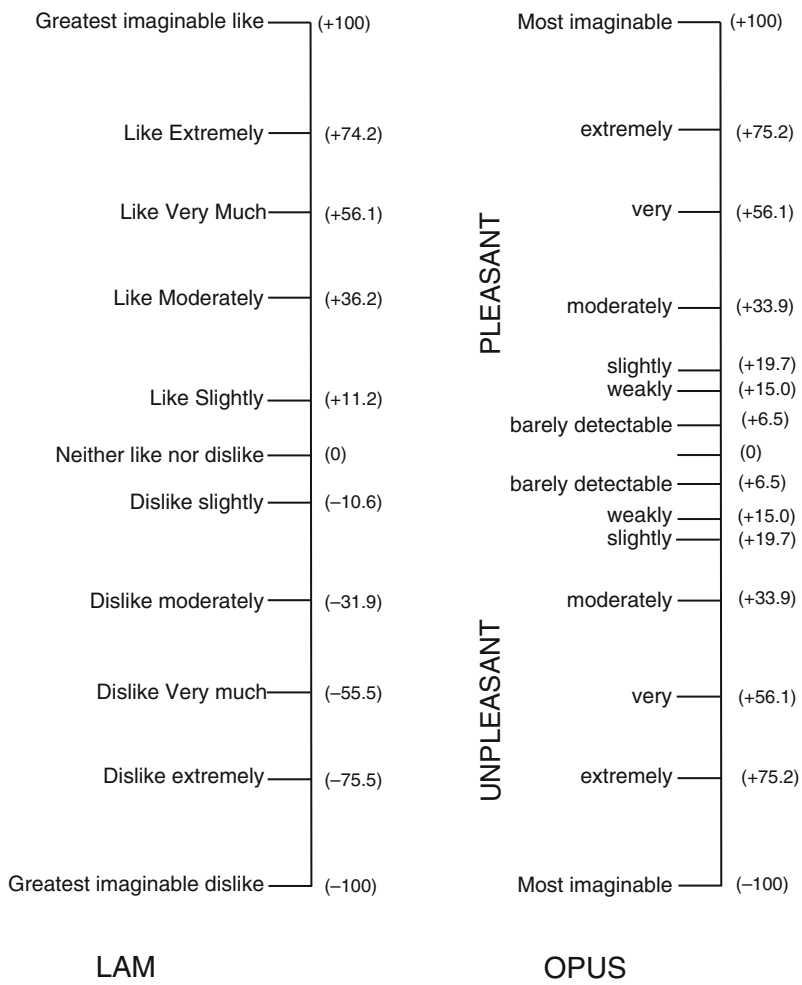


**Fig. 7.5** Affective labeled magnitude scales, including the LAM scale (Cardello and Schutz, 2004) and the OPUS scale (Guest et al., 2007).

scale) will also show such range effects. A compressed range of responses is obtained when the frame of reference is greatest imaginable like (dislike) for an "experience of any kind" rather than something more delimited like "foods and beverages." Apparently the compression is not very detrimental to the ability of the LAM scale to differentiate products (Cardello et al., 2008, but see also Lawless et al., 2010a).

Is this a suitable method for cross-subject comparisons? To the extent that Borg's assumptions of common perceptual range and the similarity of the high end-anchor experience among people are true, the method might provide one approach to valid comparisons of the ratings among different respondents. This would facilitate comparisons of clinical groups or patients with sensory disorders or genetically different individuals such as anosmics, PTC/PROP taster groups. Bartoshuk and colleagues (1999, 2003, 2004a, b, 2006) have argued that the labeled magnitude scales should anchor their endpoints to "sensations of any kind" as such a reference experience would allow different individuals to use the scale in similar ways/and thus facilitate inter-individual comparisons. Scales with this kind of high end anchor have been termed "generalized" labeled magnitude scales (or gLMS). However, the sensory evaluation practitioner should be aware of the compression effects that can occur with this kind of scale, which could potentially lead to lessened differentiation among products.

### 7.5.3 Adjustable Rating Techniques: Relative Scaling

A few methods have been tried that allow consumers or panelists to change their ratings. An example is

the "rank-rating" technique developed by O'Mahony, Kim, and colleagues (Kim and O'Mahony, 1998; Lee et al., 2001; O'Mahony et al., 2004; Park et al., 2004). In this method, the consumer has a visual scale in front of him or her on the table and after tasting a sample is instructed to physically place the sample on the scale. Subsequent samples are also tasted and placed and the important feature is that the consumer can change the position of any previous item based on their perception of the new sample(s). This procedure has relatively little to do with rank ordering per se (in fact ties can be allowed). Cordinnier and Delwiche (2008) chose the more descriptive name of "positional relative rating" for this technique.

O'Mahony and Kim examined the efficiency of this technique, mostly in simple salt solutions, on the basis of people's ordering of salt solutions in increasing concentrations (Kim and O'Mahony, 1998; Park et al., 2004). The important data were "reversals" in which a higher concentration was rated lower than a lower one, and vice versa. A good scale would minimize the amount of reversals. Given this criterion, the rank-rating has fewer errors than non-adjustable ratings. At this point it is not clear whether this apparent advantage arises because people are allowed to re-taste, or that they are allowed to re-position previously tasted items. Both factors may be important. There is some evidence that adjustable ratings can produce statistically significant ratings with fewer subjects, but the procedure can take up to twice as long as normal ratings. A limitation of this technique is that only one attribute can be evaluated at a time. If a second or third attribute is needed, the procedure starts over. This may be acceptable for consumer like/dislike ratings (O'Mahony et al., 2004), but would not be suitable for a descriptive analysis task. The option to change previous ratings is an interesting notion and is allowed by some of the sensory data-collection software packages at this time. Whether the option is advantageous should be the subject of further study. Re-tasting (Koo et al., 2002; Lee et al., 2001) is a potentially important feature of this method, and should be evaluated separately, and in consideration of the extra adaptation, fatigue, or carryover that could occur with some products.

A completely relative rating procedure is the method of Gay and Mead (Gay and Mead, 1992; Mead and Gay, 1995). In this task, a panelist inspects the entire set of samples and places the most intense (or most liked) at the top end of the scale and the least intense (or least liked) at the bottom. All other samples are distributed along the full scale. This can provide good differentiation of the products, but obviously any absolute information about what is weak or strong, liked or disliked, is lost (as is the case with magnitude estimation). Because all the ratings are truly relative, contextual effects such as contrast might be expected to be larger with such a technique, but this is not known.

Another relative scaling method is when panelists purposely rate each sample relative to a reference item, which is usually marked on the center of the response scale. Relative-to-reference scaling was studied by Larson-Powers and Pangborn (1978) and compared to traditional scaling for the descriptive profiling of beverages and gelatins. Significant differences were found with the relative scale ("anchored" in their terms) in 22.8% of all possible comparisons as compared with 19.5% of comparisons using the unanchored scale. However, panelists were given the relative scale first, and more practice using that scale. In an extensive study of both trained and untrained respondents, Stoer and Lawless (1993) found a similar advantage, with 33% of all possible comparisons were significant for the relative scaling versus 27% for traditional scaling. However, this was not a statistically significant increase based on meta-analytic comparisons (Rosenthal, 1987). The relative-to-reference scale was also discussed by Land and Shepard (1984), who noted that it facilitates comparisons across occasions that would be otherwise difficult to make. They also warn that the choice of standard may have an effect on the scaling functions that result. The task is certainly easy for subjects and was touted by Land and Shepard as showing "good reproducibility" (see also Mahoney et al., 1957). Whether the method offers any consistent advantage over traditional scaling is questionable. It may be useful in situations where comparison to a reference is a natural feature or explicit objective of the experiment at hand, for example in quality control or shelf life studies where an identified control sample is used as a baseline for comparison.

### 7.5.4 Ranking

Another alternative to traditional scaling is the use of ranking procedures. Ranking is simply ordering the products from weakest to strongest on the stated

attribute or from least liked to most liked for consumer acceptance testing. Ranking has the advantages of simplicity in instructions to subjects, simplicity in data handling, and minimal assumptions about level of measurement since the data are treated as ordinal. Although ranking tests are most often applied to hedonic data, they are also applicable to questions of sensory intensity. When asked to rank items for the intensity of a specific attribute, e.g., the sourness of taste of several fruit juices, for example, the ranking test is merely an extension of a paired comparison procedure into more than two products. Due to its simplicity, ranking may be an appropriate choice in situations where participants would have difficulty understanding scaling instructions. In working with non-literates, young children, across cultural boundaries or in linguistically challenging situations, ranking is worth considering (Coetzee and Taylor, 1996). This is especially true if decisions are likely to be very simple, e.g., whether or not two juice samples differ in perceived sourness. Ranking may allow differentiation of products that are all similar in acceptability. With medications, for example, all formulas may be to some degree unpalatable. It might be useful then to use ranking in choosing alternative flavorings in order to find the least offensive.

Analysis of ranked data is straightforward. Simple rank-sum statistics can be found in the tables published by Basker (1988) and Newell and MacFarlane (1987, see also Table J). Another very sensitive test of differences in ranked data is the Friedman test, also known as the "analysis of variance on ranks." These are discussed in Appendix B. The tests are rapid, straightforward, and easy to perform. It is also possible to convert other data to rankings. This is a conservative approach if the interval nature of the data is in question or when violations of statistical assumptions such as the normality of the data are suspect. For example, Pokorńy et al. (1986) used ranking analysis of line-marking data to compare the profiles of different raspberry beverages sweetened with aspartame.

### 7.5.5 Indirect Scales

A conservative approach to scaling is to use the variance in the data as units of measurement, rather than the numbers taken at face value. For example, we could ask how many standard deviations apart are the mean values for two products. This is a different approach to measurement than simply asking how many scale units separate the means on the response scale. On a 9-point scale, one product may receive mean rating of seven, and another nine, making the difference two scale units. If the pooled standard deviation is two units, however, they would only be one unit apart on a variability-based scale. As one example, Conner and Booth (1988) used both the slope and the variance of functions from just-right scales to derive a "tolerance discrimination ratio." This ratio represents a measure of the degree of difference along a physical continuum (such as concentration of sugar in lime drink) that observers find to make a meaningful change in their ratings of difference-from-ideal (or just-right). This is analogous to finding the size of a just-noticeable difference, but translated into the realm of hedonic scaling. Their insight was that it is not only the slope of the line that is important in determining this tolerance or liking-discrimination function, but also the variance around that function.

Variability-based scales are the basis for scaling in Thurstone's models for comparative judgment (Thurstone, 1927) and its extension into determining distances between category boundaries (Edwards, 1952). Since the scale values can be found from choice experiments as well as rating experiments, the technique is quite flexible. How this type of scaling can be applied to rating data is discussed below in the derivation of the 9-point hedonic scale words. When the scale values are derived from a choice method like the triangle test or paired comparison method, this is sometimes called "indirect scaling" (Baird and Noma, 1978; Jones, 1974). The basic operation in Thurstonian scaling of choice data is to convert the proportion correct in a choice experiment (or simply the proportions in a two-tailed test like paired preference) to *Z*-scores. The exact derivation depends upon the type of test (e.g., triangle versus 3-AFC) and the cognitive strategy used by the subject. Tables for Thurstonian scale values from various tests such as the triangle test were given by Frijters et al. (1980) and Bi (2006) and some tables are given in the Appendix. Mathematical details of Thurstonian scaling are discussed in Chapter 5.

Deriving measures of sensory differences in such indirect ways presents several problems in applied sensory evaluation so the method has not been widely used. The first problem is one of economy in data

collection. Each difference score is derived from a separate discrimination experiment such as a paired comparison test. Thus many subjects must be tested to get a good estimate of the proportion of choice, and this yields just one scale value. In direct scaling, each participant gives at least one data point for each item tasted. Direct scaling allows for easy comparisons among multiple products, while the discrimination test must be done on one pair at a time. Thus the methods of indirect scaling are not cost-efficient.

A second problem can occur if the products are too clearly different on the attribute in question, because then the proportion correct will approach 100%. At that point the scale value is undefined as they are some unknown number of standard deviations apart. So the method only works when there are small differences and some confusability of the items. In a study of many products, however, it is sometimes possible to compare only adjacent or similar items, e.g., products that differ in small degrees of some ingredient or process variable. This approach was taken by Yamaguchi (1967) in examining the synergistic taste combination of monosodium glutamate and disodium 5′ inosinate. Many different levels of the two ingredients were tasted, but because the differences between some levels were quite apparent, an incomplete design was used in which only three adjacent levels were compared.

Other applications have also used this notion of variability as a yardstick for sensory difference or sensation intensity. The original approach of Fechner in constructing a psychophysical function was to accumulate the difference thresholds or just-noticeable differences (JNDs) in order to construct the log function of psychophysical sensory intensity (Boring, 1942; Jones, 1974). McBride (1983a, b) examined whether JND-based scales might give similar results to category scales for taste intensity. Both types of scaling yielded similar results, perhaps not surprising since both tend to conform to log functions. In a study of children's preferences for different odors Engen (1974) used a paired preference paradigm, which was well suited to the abilities of young children to respond in a judgment task. He then converted the paired preference proportions to Thurstonian scale values via *Z*-scores and was able to show that the hedonic range of children was smaller than that of adults.

Another example of choice data that can be converted to scale values is best–worst scaling, in which a consumer is asked to choose the best liked and least liked samples from a set of three or more items (Jaeger et al., 2008). With three products, it can be considered a form of a ranking task. When applied to sensory intensity, this is sometimes known as maximum-difference or "max-diff." Best–worst scaling is also discussed in Section 13.7. Simple difference scores may be calculated based on the number of times an item is called best versus worst and these scores are supposed to have interval properties. If a multinomial logistic regression is performed on the data, they are theorized to have true ratio properties (Finn and Louviere, 1992). A practical problem with the method, however, is that so many products must be tasted and compared, rendering it difficult to perform with foods (Jaeger and Cardello, 2009).

The sensory professional should bear in mind that in spite of their theoretical sophistication, the indirect methods are based on variability as the main determinant of degree of difference. Thus any influence, which increases variability, will tend to decrease the measured differences among products. In the well-controlled psychophysical experiment under constant standard conditions across sessions and days, this may not be important—the primary variability lies in the resolving power of the participant (and secondarily in the sample products). But in drawing conclusions across different days, batches, panels, factories, and such, one has a less pure situation to consider. Whether one considers the Thurstonian-type indirect measures comparable across different conditions depends upon the control of extraneous variation.

## 7.6 Comparing Methods: What is a Good Scale?

A large number of empirical studies have been conducted comparing the results using different scaling methods (e.g., Birnbaum, 1982; Giovanni and Pangborn, 1983; Hein et al., 2008; Jaeger and Cardello, 2009; Lawless and Malone, 1986a, b; Lawless et al., 2010a; Marks et al., 1983; Moskowitz and Sidel, 1971; Pearce et al., 1986; Piggot and Harper, 1975; Shand et al., 1985; Vickers, 1983; Villanueva and Da Silva, 2009; Villanueva et al., 2005). Because scaling data are often used to identify differences between products, the ability to detect differences is one important

practical criterion for how useful a scaling method can be (Moskowitz and Sidel, 1971). A related criterion is the degree of error variance or similar measures such as size of standard deviations or coefficients of variation. Obviously, a scaling method with low inter-individual variability will result in more sensitive tests, more significant differences, and lower risk of Type II error (missing a true difference). A related issue is the reliability of the procedure. Similar results should be obtained upon repeated experimentation.

Other practical considerations are important as well. The task should be user friendly and easy to understand for all participants. Ideally, the method should be applicable to a wide range of products and questions, so that the respondent is not confused by changes in response type over a long ballot or questionnaire. If panelists are familiar with one scale type and are using it effectively, there may be some liability in trying to introduce a new or unfamiliar method. Some methods, like category scales, line scales, and magnitude estimation, can be applied to both intensity and hedonic (like–dislike) responses. The amount of time required to code, tabulate and process the information may be a concern, depending upon computer-assisted data collection and other resources available to the experimenters.

As in any method, validity or accuracy are also issues. Validity can only be judged by reference to some external criterion. For hedonic scaling, one might want the method to correspond to other behaviors such as choice or consumption (Lawless et al., 2010a). A related criterion is the ability of the scale to identify or uncover consumer segments with different preferences (Villanueva and Da Silva, 2009).

Given these practical considerations, we may then ask how the different scaling methods fare. Most published studies have found about equal sensitivity for the different scaling methods, provided that the methods are applied in a reasonable manner. For example, Lawless and Malone (1986a, b) performed an extensive series of studies (over 20,000 judgments) with consumers in central location tests using different sensory continua including olfaction, tactile, and visual modalities. They compared line scales, magnitude estimation, and category scales. Using the degree of statistical differentiation among products as the criterion for utility of the methods, the scales performed about equally well. A similar conclusion was reached by Shand et al. (1985) for trained panelists. There was some small tendency for magnitude estimation to be marginally more variable in the hands of consumers as opposed to college students (Lawless and Malone, 1986b). Statistical differentiation increased over replicates, as would be expected as people came to understand the range of items to be judged (see Hein et al., 2008, for another example of improvement over replication in hedonic scaling). Similar findings for magnitude estimation and category scales in terms of product differentiation were found by Moskowitz and Sidel (1971), Pearce et al. (1986), Shand et al. (1985), and Vickers (1983) although the forms of the mathematical relations to underlying physical variables was often different (Piggot and Harper, 1975). In other words, as found by Stevens and Galanter (1957) there is often a curvilinear relationship between the data from the two methods. However, this has not been universally observed and sometimes simple linear relationships have been found (Vickers, 1983). Similar results for category scales and line scales were found by Mattes and Lawless (1985).

Taken together, these empirical studies paint a picture of much more parity among methods than one might suppose given the number of arguments over the validity of scaling methods in the psychological literature. With reasonable spacing of the products and some familiarization with the range to be expected, respondents will distribute their judgments across the available scale range and use the scale appropriately to differentiate the products. A reasonable summary of the literature comparing scale types is that they work about equally well to differentiate products, given a few sensible precautions.

## 7.7 Issues

### 7.7.1 *"Do People Make Relative Judgments" Should They See Their Previous Ratings?*

Baten's (1946) report of an advantage for the line scale is illustrative of how observant many researchers were in the older literature. He noted that a category scale with labeled alternatives might help some judges, but could hinder others by limiting the alternative

responses (i.e., judgments that might fall in-between categories). The line scale offers a continuously graded choice of alternatives, limited only by the measurement abilities in data tabulation. Baten also noted that the line scale seemed to facilitate a relative comparison among the products. This was probably due to his placement of the scales one above the other on the ballot, so judges could see both markings at the same time. In order to minimize such contextual effects it is now more common to remove the prior ratings for products to achieve a more independent judgment of the products. However, whether that is ever achieved in practice is open to question—humans are naturally comparative when asked to evaluate items, as discussed in Chapter 9. Furthermore, there may be potential for increased discrimination in methods like the relative positioning technique. The naturally comparative nature of human judgment may be something we could benefit from rather than trying to fight this tendency by over-calibration.

### 7.7.2 Should Category Rating Scales Be Assigned Integer Numbers in Data Tabulation? Are They Interval Scales?

There is also a strong suspicion that many numerical scaling methods may produce only ordinal data, because the spacing between alternatives is not subjectively equal. A good example is the common marketing research scale of "excellent—very good—good—fair—poor." The subjective spacing between these adjectives is quite uneven. The difference between two products rated good and very good is a much smaller difference than that between products rated fair and poor. However, in analysis we are often tempted to assigned numbers one through five to these categories and take means and perform statistics as if the assigned numbers reflected equal spacing. This is a pretense at best. A reasonable analysis of the 5-point excellent to poor scale is simply to count the number of respondents in each category and to compare frequencies. Sensory scientists should not assume that any scale has interval properties in spite of how easy it is to tabulate data as an integer series.

### 7.7.3 Is Magnitude Estimation a Ratio Scale or Simply a Scale with Ratio Instructions?

In magnitude estimation, subjects are instructed to use numbers to reflect the relative proportions between the intensities experienced from different stimuli. A beverage that is twice as sweet as another should be given a response number that is twice as large. S. S. Stevens felt that these numbers were accurate reflections of the experience, and so the scale had ratio properties. This assumed a linear relationship between the subjective stimulus intensities (sensations or percepts) and the numerical responses. However, there is a wealth of information showing that the process of numerical assignment is prone to a series of contextual and number usage biases that strongly question whether this process is linear (Poulton, 1989). So Stevens' original view of accepting the numerical reporting as having face validity seems misplaced. Although it would be advantageous to achieve a level of measurement that did allow conclusions about proportions and ratios ("this is liked twice as much as that"), this seems not fully justified at this time. It is important to differentiate between a method that has ratio-type instructions, and one that yields a true ratio scale of sensation magnitude, where the numbers actually reflect proportions between intensities of sensory experiences.

### 7.7.4 What is a "Valid" Scale?

An ongoing issue in psychophysics is what kind of scale is a true reflection of the subject's actual sensations? From this perspective, a scale is valid when the numbers generated reflect a linear translation of subjective intensity (the private experience). It is well established that category scales and magnitude estimates, when given to the same stimuli, will form a curve when plotted against one another (Stevens and Galanter, 1957). Because this is not a linear relationship, one method or the other must result from a non-linear translation of the subjective intensities of the stimuli. Therefore, by this criterion, at least one scale must be "invalid."

Anderson (1974, 1977) proposed a functional measurement theory to address this issue. In a typical experiment, he would ask subjects to do some kind of combination task, like judging the total combined intensity of two separately presented items (or the average lightness of two gray swatches). He would set up a factorial design in which every level of one stimulus was combined with every level of the other (i.e., a complete block). When plotting the response, a family of lines would be seen when the first stimulus continuum formed the *X*-axis and the second formed a family of lines. Anderson argued that only when the response combination rule was additive, *and* the response output function was linear, would a parallel plot be obtained (i.e., there would be no significant interaction term in ANOVA). This argument is illustrated in Fig. 7.6. In his studies using simple line and category scales, parallelism was obtained in a number of studies, and thus he reasoned that magnitude estimation was invalid by this criterion. If magnitude estimation is invalid, then its derivatives such as the LMS and LAM scales are similarly suspect.

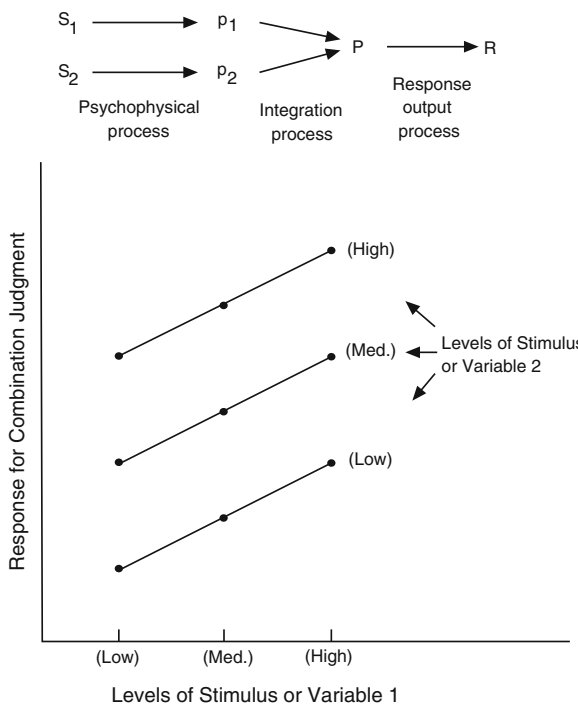


**Fig. 7.6** The functional measurement scheme of Anderson (1974).

Others have found support for the validity of magnitude estimation in studies of binaural loudness summation (Marks, 1978). This argument continues and is difficult to resolve. A review of the matter was published by Gescheider (1988). For the purposes of sensory evaluation, the issue is not terribly important for two reasons. First, any scale that produces statistically significant differentiation of products is a useful scale. Second, the physical ranges over which category scaling and magnitude estimation produce different results is usually quite large in any psychophysical study. In most product tests, the differences are much more subtle and generally do not span such a wide dynamic range. The issue dissolves from any practical perspective.

## 7.8 Conclusions

Much sound and fury has been generated over the years in the psychophysical literature concerning what methods yield valid scales. For the sensory practitioner, these issues are less relevant because the scale values do not generally have any absolute meaning. They are only convenient indices of the relative intensities or appeal of different products. The degree of difference may be a useful piece of information, but often we are simply interested in which product is stronger or weaker in some attribute, and whether the difference is both statistically significant and practically meaningful.

Scaling provides a quick and useful way to get intensity or liking information. In the case of descriptive analysis, scaling allows collection of quantitative data on multiple attributes. The degree of variability or noise in the system is, to a large part, determined by whether the panelists have a common frame of reference. Thus reference standards for both the attribute terms and for the intensity anchors are useful. Of course, with consumer evaluations or a psychophysical study such calibration is not possible and usually not desired. The variability of consumer responses should offer a note of caution in the interpretation of consumer scaling data.

Students and sensory practitioners should examine their scaling methods with a critical eye. Not every task that assigns numbers will have useful scale properties like equal intervals. Bad examples abound

in the commodity grading (quality scoring) literature (Pangborn and Dunkley, 1964). For example, different numbers may be assigned to different levels of oxidation, but that is scoring a physical condition based on inferences from sensory experience. It is not a report of the intensity of some experience itself. It is not tracking changes along a single perceptual continuum in the psychophysical sense. Scoring is not scaling.

> All hedonic scales seem to measure what they are intended to measure rather effectively, as long as no gross mistakes are made (Peryam, 1989, p. 23).

## Appendix 1: Derivation of Thurstonian-Scale Values for the 9-Point Scale

The choice of adjective words for the 9-point hedonic scale is a good example of how carefully a scale can be constructed. The long-standing track record of this tool demonstrates its utility and wide applicability in consumer testing. However, few sensory practitioners actually know how the adjectives were found and what criteria were brought to bear in selecting these descriptors (slightly, moderately, very much, and extremely like/dislike) from a larger pool of possible words. The goal of this section is to provide a shorthand description of the criteria and mathematical method used to select the words for this scale.

One concern was the degree to which the term had consensual meaning in the population. The most serious concern was when a candidate word had an ambiguous or double meaning across the population. For example, the word "average" suggests an intermediate response to some people, but in the original study by Jones and Thurstone (1955) there were a group of people who equated it with "like moderately" perhaps since an average product in those days was one that people would like. These days, one can think of negative connotations to the word "average" as in "he was only an average student." Other ambiguous or bimodal terms were "like not so much" and "like not so well." Ideally, a term should have low variability in meaning, i.e., a low standard deviation, no bimodality, and little skew. Part of this concern with the normality of the distribution of psychological reactions to a word was the fact that the developers used Thurstone's model for categorical judgment as a means of measuring the psychological-scale values for the words. This model is at its most simple form when the items to be scaled show normal distributions of equal variance.

Which leads us to the numerical method. Jones and Thurstone modified a procedure used earlier by Edwards (1952). A description of the process and results can be found in the paper "Development of a scale for measuring soldiers' food preferences" by Jones et al. (1955). Fifty-one words and phrases formed the candidate list based on a pilot study with 900 soldiers chosen to be a representative sample of enlisted personnel. Each phrase was presented on a form with a rating scale from –4 to +4 with a check off format. In other words, each person read each phrase and assigned in an integer value from –4 to +4 (including zero as an option). This method would seem to presume that these integers were themselves an interval scale of psychological magnitude, an assumption that to our knowledge has never been questioned.

Of course, the mean scale values could now be assigned on a simple and direct basis, but the Thurstonian methods do not use the raw numbers as the scale, but transform them to use standard deviations as the units of measurement. So the scale needs to be converted to *Z*-score values. The exact steps are as follows:

1. Accumulate frequency counts for all the tested words across the –4 to + 4 scale. Think of these categories as little "buckets" into which judgments have been tossed.
2. Find the marginal proportions each value from –4 to +4 (summed across all test items). Add up the proportions from lowest to highest to get a cumulative proportion for each bucket.
3. Convert these proportions to *z*-scores in order to re-scale the boundaries for the original −4 to +4 cutoffs. Let us call these the "category *z*-values" for each of the "buckets." The top bucket will have a value of 100%, so it will have no *z*-score (undefined/infinite).
4. Next examine each individual item. Sum its individual proportions across the categories, from where it is first used until 100% of the responses are accumulated.
5. Convert the proportions for the item to *Z*-scores. Alternatively, you can plot these proportions on "cumulative probability paper," a graphing format

that marks the ordinate in equal standard deviations units according to the cumulative normal distribution. Either of these methods will tend to make the cumulative S-shaped curve for the item into a straight line. The *X*-axis value for each point is the "category *z*-value" for that bucket.

6. Fit a line to the data and interpolate the 50% point on the *X*-axis (the re-scaled category boundary estimates). These interpolated values for the median for each item now form the new scale values for the items.

An example of this interpolation is shown in Fig. 7.7. Three of the phrases used in the original scaling study of Jones and Thurstone (1955) are pictured, three that were not actually chosen but for which we have approximate proportions and *z*-scores from their figures. The small vertical arrows on the *X*-axis show the scale values for the original categories of −4 to +3 (+4 has cumulative proportion of 100% and thus the *z*-score is infinite). Table 7.1 gives the values and proportions for each phrase and the original categories. The dashed vertical lines dropped from the intersection at the zero *z*-score (50% point) show the approximate mean values interpolated on the *X*-axis (i.e., about –1.1 for "do not care for it" and about +2.1 for "preferred."). Note that "preferred" and "don't care for it" have a linear fit and steep slope, suggesting a normal distribution and low standard deviation. In contrast, "highly unfavorable" has a lower slope and some curvilinearity, indicative of higher variability, skew, and/or pockets of disagreement about the connotation of this term.

The actual scale values for the original adjectives are shown in Table 7.2, as found with a soldier population circa 1950 (Jones et al., 1955). You may note that the words are not equally spaced, and that the "slightly" values are closer to the neutral point than some of the other intervals, and the extreme points are a little farther out. This bears a good deal of similarity to the intervals found with the LAM scale as shown in the column where the LAM values are re-scaled to the same range as the 9-point Thurstonian Values.

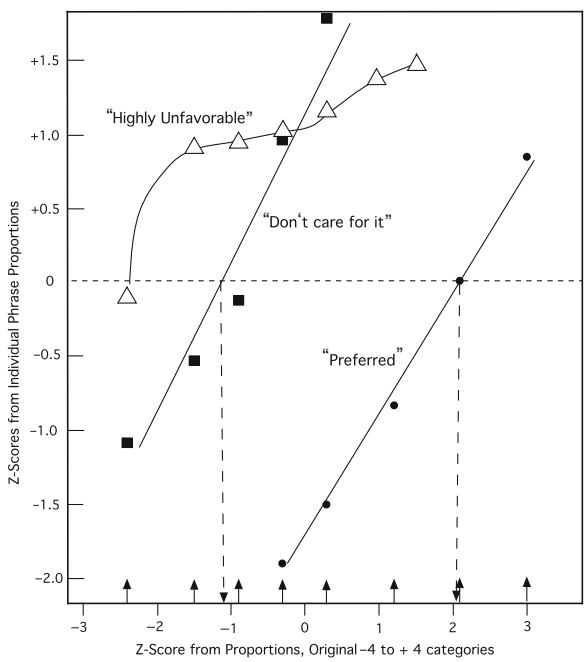


**Fig. 7.7** An illustration of the method used to establish spacings and scale values for the 9-point hedonic scale using Thurstonian theory. *Arrows* on the *X*-axis show the scale points for the *z*-scores based on the complete distribution of the original –4 to +4 ratings. The *Y*-axis shows the actual *z*-scores based on the proportion of respondents using that category for each specific term. Re-plotted from data provided in Jones et al. (1955).

## Appendix 2: Construction of Labeled Magnitude Scales

There are two primary methods for constructing labeled magnitude scales and they are very similar. Both require magnitude estimates from the participants to scale the word phrases used on the lines. In one case, just the word phrases are scaled, and in the second method, the word phrases are scaled among a list of common everyday experiences or sensations that most people are familiar with. The values obtained by the simple scaling of just the words will depend upon the words that are chosen, and extremely high examples (e.g., greatest imaginable liking for any experience) will tend to compress the values of the interior phrases (Cardello et al., 2008). Whether this kind of context effect will occur for the more general method of scaling amongst common experiences is not known. But the use of a broad frame of reference could be a stabilizing factor.

Here is an example of the instructions given to subjects in construction of a labeled affective magnitude scale. Note that for hedonics, which are a bipolar continuum with a neutral point, it is necessary to collect a tone or valence (plus or minus) value as well as the overall "intensity" rating.

**Table 7.1** Examples of scaled phrases used in Fig. 7.7

| Original | | | "Preferred" | | "Do not care for it" | | "Highly unfavorable" | |
|---|---|---|---|---|---|---|---|---|
| category | Proportion | *Z*-score | Proportion | *Z*-score | Proportion | *Z*-score | Proportion | Z-score |
| 4 | 1.000 | (undef.) | 0.80 | 0.84 | | | | |
| 3 | 0.999 | 3.0 | 0.50 | 0.00 | | | 0.96 | 1.75 |
| 2 | 0.983 | 2.1 | 0.20 | −0.84 | | | 0.93 | 1.48 |
| 1 | 0.882 | 1.2 | 0.07 | −1.48 | | | 0.92 | 1.41 |
| 0 | 0.616 | 0.3 | 0.03 | −1.88 | 0.96 | 1.75 | 0.90 | 1.28 |
| −1 | 0.383 | −0.3 | | | 0.83 | 0.95 | 0.86 | 1.08 |
| −2 | 0.185 | −0.9 | | | 0.55 | 0.13 | 0.84 | 0.99 |
| −3 | 0.068 | −1.5 | | | 0.30 | −0.52 | 0.82 | 0.92 |
| −4 | 0.008 | −2.4 | | | 0.14 | −1.08 | 0.46 | −0.10 |

**Table 7.2** Actual 9-point scale phrase values and comparison to the LAM values

| Descriptor | Scale value (9-point) | Interval | LAM value | LAM rescaled | Interval |
|---|---|---|---|---|---|
| Like extremely | 4.16 | | 74.2 | 4.20 | |
| Like very much | 2.91 | 1.26 | 56.1 | 3.18 | 1.02 |
| Like moderately | 1.12 | 1.79 | 36.2 | 2.05 | 1.13 |
| Like slightly | 0.69 | 0.43 | 11.2 | 0.63 | 1.52 |
| Neither like nor dislike | 0.00 | 0.69 | 0.0 | 0.00 | 0.63 |
| Dislike slightly | −0.59 | 0.59 | −10.6 | −0.60 | 0.60 |
| Dislike moderately | −1.20 | 0.61 | −31.9 | −1.81 | 1.21 |
| Dislike very much | −2.49 | 1.29 | −55.5 | −3.14 | 1.33 |
| Dislike extremely | −4.32 | 1.83 | −75.5 | −4.28 | 1.14 |

Next to each word label a response area appeared similar to this:

Phrase: Tone: + − 0 How much:
Like extremely _____ _____

Words or phrases are presented in random order. After reading a word they must decide whether the word is positive, negative or neutral and place the corresponding symbol on the first line. If the hedonic tone was not a neutral one (zero value), they are instructed to give a numerical estimate using modulus-free magnitude estimation. The following is a sample of the instructions taken from Cardello et al. (2008):

> After having determined whether the phrase is positive or negative or neutral and writing the appropriate symbol (+, −, 0) on the first line, you will then assess the strength or magnitude of the liking or disliking reflected by the phrase. You will do this by placing a number on the second blank line (under "How Much"). For the first phrase that you rate, you can write any number you want on the line. We suggest you do not use a small number for this word/phrase. The reason for this is that subsequent words/phrases may reflect much lower levels of liking or disliking. Aside from this restriction you can use any numbers you want. For each subsequent word/phrase your numerical judgment should be made proportionally and in comparison to the first number. That is, if you assigned the number 800 to index the strength of the liking/disliking denoted by the first word/phrase and the strength of liking/disliking denoted by the second word/phrase were twice as great, you would assign the number 1,600. If it were three times as great you would assign the number 2,400, etc. Similarly, if the second word/phrase denoted only 1/10 the magnitude of liking as the first, you would assign it the number 80 and so forth. If any word/phrase is judged to be "neutral" (zero (0) on the first line) it should also be given a zero for its magnitude rating.

In the cased of Cardello et al. (2008), positive and negative word labels were analyzed separately. Raw magnitude estimates were equalized for scale range using the procedure of Lane et al. (1961). All positive and negative magnitude estimates for a given subject were multiplied by an individual scaling factor. This factor was equal to the ratio of the grand geometric mean (of the absolute value of all nonzero ratings) across all subjects divided by the geometric mean for that subject. The geometric mean magnitude estimates for each phrase were then calculated based on this range-equated data. These means became the distance

from the zero point for placement of the phrases along the scale, usually accompanied by a short cross-hatch mark at that point.

## References


AACC (American Association of Cereal Chemists). 1986. Approved Methods of the AACC, Eighth Edition. Method 90–10. Baking quality of cake flour, rev. Oct. 1982. The American Association of Cereal Chemists, St. Paul, MN, pp. 1–4.

Anderson, N. H. 1974. Algebraic models in perception. In: E. C. Carterette and M. P. Friedman (eds.), Handbook of Perception. Psychophysical Judgment and Measurement, Vol. 2. Academic, New York, pp. 215–298.

Anderson, N. H. 1977. Note on functional measurement and data analysis. Perception and Psychophysics, 21, 201–215.

ASTM. 2008a. Standard test method for unipolar magnitude estimation of sensory attributes. Designation E 1697-05. In: Annual Book of ASTM Standards, Vol. 15.08, End Use Products. American Society for Testing and Materials, Conshohocken, PA, pp. 122–131.

ASTM. 2008b. Standard test method for sensory evaluation of red pepper heat. Designation E 1083-00. In: Annual Book of ASTM Standards, Vol. 15.08, End Use Products. American Society for Testing and Materials, Conshohocken, PA, pp. 49–53.

Aust, L. B., Gacula, M. C., Beard, S. A. and Washam, R. W., II. 1985. Degree of difference test method in sensory evaluation of heterogeneous product types. Journal of Food Science, 50, 511–513.

Baird, J. C. and Noma, E. 1978. Fundamentals of Scaling and Psychophysics. Wiley, New York.

Banks, W. P. and Coleman, M. J. 1981. Two subjective scales of number. Perception and Psychophysics, 29, 95–105.

Bartoshuk, L. M., Snyder, D. J. and Duffy, V. B. 2006. Hedonic gLMS: Valid comparisons for food liking/disliking across obesity, age, sex and PROP status. Paper presented at the 2006 Annual Meeting, Association for Chemoreception Sciences.

Bartoshuk, L. M., Duffy, V. B., Fast, K., Green, B. G., Prutkin, J. and Snyder, D. J. 2003. Labeled scales (e.g. category, Likert, VAS) and invalid across-group comparisons: What we have learned from genetic variation in taste. Food Quality and Preference, 14, 125–138.

Bartoshuk, L. M., Duffy, V. B., Green, B. G., Hoffman, H. J., Ko, C.-W., Lucchina, L. A., Marks, L. E., Snyder, D. J. and Weiffenbach, J. M. 2004a. Valid across-group comparisons with labeled scales: the gLMS versus magnitude matching. Physiology and Behavior, 82, 109–114.

Bartoshuk, L. M., Duffy, V. B., Chapo, A. K., Fast, K., Yiee, J. H., Hoffman, H. J., Ko, C.-W. and Snyder, D. J. 2004b. From psychophysics to the clinic: Missteps and advances. Food Quality and Preference, 14, 617–632.

Bartoshuk, L. M., Duffy, V. B., Fast, K., Green, B. Kveton, J., Lucchina, L. A., Prutkin, J. M., Snyder, D. J. and Tie, K. 1999. Sensory variability, food preferences and BMI in non-medium and supertasters of PROP. Appetite, 33, 228–229.

Basker, D. 1988. Critical values of differences among rank sums for multiple comparisons. Food Technology, 42(2), 79, 80–84.

Baten, W. D. 1946. Organoleptic tests pertaining to apples and pears. Food Research, 11, 84–94.

Bendig, A. W. and Hughes, J. B. 1953. Effect of number of verbal anchoring and number of rating scale categories upon transmitted information. Journal of Experimental Psychology, 46(2), 87–90.

Bi, J. 2006. Sensory Discrimination Tests and Measurement. Blackwell, Ames, IA.

Birch, L. L., Zimmerman, S. I. and Hind, H. 1980. The influence of social-affective context on the formation of children's food preferences. Child Development, 51, 865–861.

Birch, L. L., Birch, D., Marlin, D. W. and Kramer, L. 1982. Effects of instrumental consumption on children's food preferences. Appetite, 3, 125–143.

Birnbaum, M. H. 1982. Problems with so-called "direct" scaling. In: J. T. Kuznicki, R. A. Johnson and A. F. Rutkiewic (eds.), Selected Sensory Methods: Problems and Approaches to Hedonics. American Society for Testing and Materials, Philadelphia, pp. 34–48.

Borg, G. 1982. A category scale with ratio properties for intermodal and interindividual comparisons. In: H.-G. Geissler and P. Pextod (Eds.), Psychophysical Judgment and the Process of Perception. VEB Deutscher Verlag der Wissenschaften, Berlin, pp. 25–34.

Borg, G. 1990. Psychophysical scaling with applications in physical work and the perception of exertion. Scandinavian Journal of Work and Environmental Health, 16, 55–58.

Boring, E. G. 1942. Sensation and Perception in the History of Experimental Psychology. Appleton-Century-Crofts, New York.

Brandt, M. A., Skinner, E. Z. and Coleman, J. A. 1963. The texture profile method. Journal of Food Science, 28, 404–409.

Butler, G., Poste, L. M., Wolynetz, M. S., Agar, V. E. and Larmond, E. 1987. Alternative analyses of magnitude estimation data. Journal of Sensory Studies, 2, 243–257.

Cardello, A. V. and Schutz, H. G. 2004. Research note. Numerical scale-point locations for constructing the LAM (Labeled affective magnitude) scale. Journal of Sensory Studies, 19, 341–346.

Cardello, A. V., Lawless, H. T. and Schutz, H. G. 2008. Effects of extreme anchors and interior label spacing on labeled magnitude scales. Food Quality and Preference, 21, 323–334.

Cardello, A. V., Winterhaler, C. and Schutz, H. G. 2003. Predicting the handle and comfort of military clothing fabrics from sensory and instrumental data: Development and application of new psychophysical methods. Textile Research Journal, 73, 221–237.

Cardello, A. V., Schutz, H. G., Lesher, L. L. and Merrill, E. 2005. Development and testing of a labeled magnitude scale of perceived satiety. Appetite, 44, 1–13.

Caul, J. F. 1957. The profile method of flavor analysis. Advances in Food Research, 7, 1–40.

Chambers, E. C. and Wolf, M. B. 1996. Sensory Testing Methods. ASTM Manual Series, MNL 26. ASTM International, West Conshohocken, PA.

Chen, A. W., Resurreccion, A. V. A. and Paguio, L. P. 1996. Age appropriate hedonic scales to measure the food preferences of young children. Journal of Sensory Studies, 11, 141–163.

Chung, S.-J. and Vickers, 2007a. Long-term acceptability and choice of teas differing in sweetness. Food Quality and Preference 18, 963–974.

Chung, S.-J. and Vickers, 2007b. Influence of sweetness on the sensory-specific satiety and long-term acceptability of tea. Food Quality and Preference, 18, 256–267.

Coetzee, H. and Taylor, J. R. N. 1996. The use and adaptation of the paired comparison method in the sensory evaluation of hamburger-type patties by illiterate/semi-literate consumers. Food Quality and Preference, 7, 81–85.

Collins, A. A. and Gescheider, G. A. 1989. The measurement of loudness in individual children and adults by absolute magnitude estimation and cross modality matching. Journal of the Acoustical Society of America, 85, 2012–2021.

Conner, M. T. and Booth, D. A. 1988. Preferred sweetness of a lime drink and preference for sweet over non-sweet foods. Related to sex and reported age and body weight. Appetite, 10, 25–35.

Cordinnier, S. M. and Delwiche, J. F. 2008. An alternative method for assessing liking: Positional relative rating versus the 9-point hedonic scale. Journal of Sensory Studies, 23, 284–292.

Cox, E. P. 1980. The optimal number of response alternatives for a scale: A review. Journal of Marketing Research, 18, 407–422.

Curtis, D. W., Attneave, F. and Harrington, T. L. 1968. A test of a two-stage model of magnitude estimation. Perception and Psychophysics, 3, 25–31.

Edwards, A. L. 1952. The scaling of stimuli by the method of successive intervals. Journal of Applied Psychology, 36, 118–122.

Ekman, G. 1964. Is the power law a special case of Fechner's law? Perceptual and Motor Skills, 19, 730.

Einstein, M. A. 1976. Use of linear rating scales for the evaluation of beer flavor by consumers. Journal of Food Science, 41, 383–385.

El Dine, A. N. and Olabi, A. 2009. Effect of reference foods in repeated acceptability tests: Testing familiar and novel foods using 2 acceptability scales. Journal of Food Science, 74, S97–S105.

Engen, T. 1974. Method and theory in the study of odor preferences. In: A. Turk, J. W. Johnson and D. G. Moulton (Eds.), Human Responses to Environmental Odors. Academic, New York.

Finn, A. and Louviere, J. J. 1992. Determining the appropriate response to evidence of public concern: The case of food safety. Journal of Public Policy and Marketing, 11, 12–25.

Forde, C. G. and Delahunty, C. M. 2004. Understanding the role cross-modal sensory interactions play in food acceptability in younger and older consumers. Food Quality and Preference, 15, 715–727.

Frijters, J. E. R., Kooistra, A. and Vereijken, P. F. G. 1980. Tables of d' for the triangular method and the 3-AFC signal detection procedure. Perception and Psychophysics, 27, 176–178.

Gaito, J. 1980. Measurement scales and statistics: Resurgence of an old misconception. Psychological Bulletin, 87, 564–587.

Gay, C., and Mead, R. 1992 A statistical appraisal of the problem of sensory measurement. Journal of Sensory Studies, 7, 205–228.

Gent, J. F. and Bartoshuk, L. M. 1983. Sweetness of sucrose, neohesperidin dihydrochalcone and sacchar in is related to genetic ability to taste the bitter substance 6-*n*-propylthiouracil. Chemical Senses, 7, 265–272.

Gescheider, G. A. 1988. Psychophysical scaling. Annual Review of Psychology, 39, 169–200.

Giovanni, M. E. and Pangborn, R. M. 1983. Measurement of taste intensity and degree of liking of beverages by graphic scaling and magnitude estimation. Journal of Food Science, 48, 1175–1182.

Gracely, R. H., McGrath, P. and Dubner, R. 1978a. Ratio scales of sensory and affective verbal-pain descriptors. Pain, 5, 5–18.

Gracely, R. H., McGrath, P. and Dubner, R. 1978b. Validity and sensitivity of ratio scales of sensory and affective verbal-pain descriptors: Manipulation of affect by Diazepam. Pain, 5, 19–29.

Green, B. G., Shaffer, G. S. and Gilmore, M. M. 1993. Derivation and evaluation of a semantic scale of oral sensation magnitude with apparent ratio properties. Chemical Senses, 18, 683–702.

Green, B. G., Dalton, P., Cowart, B., Shaffer, G., Rankin, K. and Higgins, J. 1996. Evaluating the "Labeled Magnitude Scale" for measuring sensations of taste and smell. Chemical Senses, 21, 323–334.

Greene, J. L., Bratka, K. J., Drake, M. A. and Sanders, T. H. 2006. Effective of category and line scales to characterize consumer perception of fruity fermented flavors in peanuts. Journal of Sensory Studies, 21, 146–154.

Guest, S., Essick, G., Patel, A., Prajpati, R. and McGlone, F. 2007. Labeled magnitude scales for oral sensations of wetness, dryness, pleasantness and unpleasantness. Food Quality and Preference, 18, 342–352.

Hein, K. A., Jaeger, S. R., Carr, B. T. and Delahunty, C. M. 2008. Comparison of five common acceptance and preference methods. Food Quality and Preference, 19, 651–661.

Huskisson, E. C. 1983. Visual analogue scales. In: R. Melzack (Ed.), Pain Measurement and Assessment. Raven, New York, pp. 34–37.

Jaeger, S. R.; Jørgensen, A. S., AAslying, M. D. and Bredie, W. L. P. 2008. Best-worst scaling: An introduction and initial comparison with monadic rating for preference elicitation with food products. Food Quality and Preference, 19, 579–588.

Jaeger, S. R. and Cardello, A. V. 2009. Direct and indirect hedonic scaling methods: A comparison of the labeled affective magnitude (LAM) scale and best-worst scaling. Food Quality and Preference, 20, 249–258.

Jones, F. N. 1974. History of psychophysics and judgment. In: E. C. Carterette and M. P. Friedman (Eds.), Handbook of Perception. Psychophysical Judgment and Measurement, Vol. 2. Academic, New York, pp. 11–22.

Jones, L. V. and Thurstone, L. L. 1955. The psychophysics of semantics: An experimental investigation. Journal of Applied Psychology, 39, 31–36.

Jones, L. V., Peryam, D. R. and Thurstone, L. L. 1955. Development of a scale for measuring soldier's food preferences. Food Research, 20, 512–520.

Keskitalo, K. Knaapila, A., Kallela, M., Palotie, A., Wessman, M., Sammalisto, S., Peltonen, L., Tuorila, H. and Perola, M. 2007. Sweet taste preference are partly genetically

determined: Identification of a trait locus on Chromosome 16[1–3]. American Journal of Clinical Nutrition, 86, 55–63.
Kim, K.-O. and O'Mahony, M. 1998. A new approach to category scales of intensity I: Traditional versus rank-rating. Journal of Sensory Studies, 13, 241–249.
King, B. M. 1986. Odor intensity measured by an audio method. Journal of Food Science, 51, 1340–1344.
Koo, T.-Y., Kim, K.-O., and O'Mahony, M. 2002. Effects of forgetting on performance on various intensity scaling protocols: Magnitude estimation and labeled magnitude scale (Green scale). Journal of Sensory Studies, 17, 177–192.
Kroll, B. J. 1990. Evaluating rating scales for sensory testing with children. Food Technology, 44(11), 78–80, 82, 84, 86.
Kurtz, D. B., White, T. L. and Hayes, M. 2000. The labeled dissimilarity scale: A metric of perceptual dissimilarity. Perception and Psychophysics, 62, 152–161.
Land, D. G. and Shepard, R. 1984. Scaling and ranking methods. In: J. R. Piggott (ed.), Sensory Analysis of Foods. Elsevier Applied Science, London, pp. 141–177.
Lane, H. L., Catania, A. C. and Stevens, S. S. 1961. Voice level: Autophonic scale, perceived loudness and effect of side tone. Journal of the Acoustical Society of America, 33, 160–167.
Larson-Powers, N. and Pangborn, R. M. 1978. Descriptive analysis of the sensory properties of beverages and gelatins containing sucrose or synthetic sweeteners. Journal of Food Science, 43, 47–51.
Lawless, H. T. 1977. The pleasantness of mixtures in taste and olfaction. Sensory Processes, 1, 227–237.
Lawless, H. T. 1989. Logarithmic transformation of magnitude estimation data and comparisons of scaling methods. Journal of Sensory Studies, 4, 75–86.
Lawless, H. T. and Clark, C. C. 1992. Psychological biases in time intensity scaling. Food Technology, 46, 81, 84–86, 90.
Lawless, H. T. and Malone, J. G. 1986a. The discriminative efficiency of common scaling methods. Journal of Sensory Studies, 1, 85–96.
Lawless, H. T. and Malone, G. J. 1986b. A comparison of scaling methods: Sensitivity, replicates and relative measurement. Journal of Sensory Studies, 1, 155–174.
Lawless, H. T. and Skinner, E. Z. 1979. The duration and perceived intensity of sucrose taste. Perception and Psychophysics, 25, 249–258.
Lawless, H. T., Popper, R. and Kroll, B. J. 2010a. Comparison of the labeled affective magnitude (LAM) scale, an 11-point category scale and the traditional nine-point hedonic scale. Food Quality and Preference, 21, 4–12.
Lawless, H. T., Sinopoli, D. and Chapman, K. W. 2010b. A comparison of the labeled affective magnitude scale and the nine point hedonic scale and examination of categorical behavior. Journal of Sensory Studies, 25, S1, 54–66.
Lawless, H. T., Cardello, A. V., Chapman, K. W., Lesher, L. L., Given, Z. and Schutz, H. G. 2010c. A comparison of the effectiveness of hedonic scales and end-anchor compression effects. Journal of Sensory Studies, 28, S1, 18–34.
Lee, H.-J., Kim, K.-O., and O'Mahony, M. 2001. Effects of forgetting on various protocols for category and line scales of intensity. Journal of Sensory Studies, 327–342.
Likert, R. 1932. Technique for the measurement of attitudes. Archives of Psychology, 140, 1–55.
Lindvall, T. and Svensson, L. T. 1974. Equal unpleasantness matching of malodourous substances in the community. Journal of Applied Psychology, 59, 264–269.
Mahoney, C. H., Stier, H. L. and Crosby, E. A. 1957. Evaluating flavor differences in canned foods. II. Fundamentals of the simplified procedure. Food Technology 11, Supplemental Symposium Proceedings, 37–42.
Marks, L. E. 1978. Binaural summation of the loudness of pure tones. Journal of the Acoustical Society of America, 64, 107–113.
Marks, L. E., Borg, G. and Ljunggren, G. 1983. Individual differences in perceived exertion assessed by two new methods. Perception and Psychophysic, 34, 280–288.
Marks, L. E., Borg, G. and Westerlund, J. 1992. Differences in taste perception assessed by magnitude matching and by category-ratio scaling. Chemical Senses, 17, 493–506.
Mattes, R. D. and Lawless, H. T. 1985. An adjustment error in optimization of taste intensity. Appetite, 6, 103–114.
McBride, R. L. 1983a. A JND-scale/category scale convergence in taste. Perception and Psychophysics, 34, 77–83.
McBride, R. L. 1983b. Taste intensity and the case of exponents greater than 1. Australian Journal of Psychology, 35, 175–184.
McBurney, D. H. and Shick, T. R. 1971. Taste and water taste for 26 compounds in man. Perception and Psychophysics, 10, 249–252.
McBurney, D. H. and Bartoshuk, L. M. 1973. Interactions between stimuli with different taste qualities. Physiology and Behavior, 10, 1101–1106.
McBurney, D. H., Smith, D. V. and Shick, T. R. 1972. Gustatory cross-adaptation: Sourness and bitterness. Perception and Psychophysics, 11, 228–232.
Mead, R. and Gay, C. 1995. Sequential design of sensory trials. Food Quality and Preference, 6, 271–280.
Mecredy, J. M. Sonnemann, J. C. and Lehmann, S. J. 1974. Sensory profiling of beer by a modified QDA method. Food Technology, 28, 36–41.
Meilgaard, M., Civille, G. V. and Carr, B. T. 2006. Sensory Evaluation Techniques, Fourth Edition. CRC, Boca Raton, FL.
Moore, L. J. and Shoemaker, C. F. 1981. Sensory textural properties of stabilized ice cream. Journal of Food Science, 46, 399–402.
Moskowitz, H. R. 1971. The sweetness and pleasantness of sugars. American Journal of Psychology, 84, 387–405.
Moskowitz, H. R. and Sidel, J. L. 1971. Magnitude and hedonic scales of food acceptability. Journal of Food Science, 36, 677–680.
Muñoz, A. M. and Civille, G. V. 1998. Universal, product and attribute specific scaling and the development of common lexicons in descriptive analysis. Journal of Sensory Studies, 13, 57–75.
Newell, G. J. and MacFarlane, J. D. 1987. Expanded tables for multiple comparison procedures in the analysis of ranked data. Journal of Food Science, 52, 1721–1725.
Olabi, A. and Lawless, H. T. 2008. Persistence of context effects with training and reference standards. Journal of Food Science, 73, S185–S189.
O'Mahony, M., Park, H., Park, J. Y. and Kim, K.-O. 2004. Comparison of the statistical analysis of hedonic data using

analysis of variance and multiple comparisons versus and R-index analysis of the ranked data. Journal of Sensory Studies, 19, 519–529.

Pangborn, R. M. and Dunkley, W. L. 1964. Laboratory procedures for evaluating the sensory properties of milk. Dairy Science Abstracts, 26–55–62.

Parducci, A. 1965. Category judgment: A range-frequency model. Psychological Review, 72, 407–418.

Park, J.-Y., Jeon, S.-Y., O'Mahony, M. and Kim, K.-O. 2004. Induction of scaling errors. Journal of Sensory Studies, 19, 261–271.

Pearce, J. H., Korth, B. and Warren, C. B. 1986. Evaluation of three scaling methods for hedonics. Journal of Sensory Studies, 1, 27–46.

Peryam. D. 1989. Reflections. In: Sensory Evaluation. In Celebration of our Beginnings. American Society for Testing and Materials, Philadelphia, pp. 21–30.

Peryam, D. R. and Girardot, N. F. 1952. Advanced taste-test method. Food Engineering, 24, 58–61, 194.

Piggot, J. R. and Harper, R. 1975. Ratio scales and category scales for odour intensity. Chemical Senses and Flavour, 1, 307–316.

Pokorńy, J., Davídek, J., Prnka, V. and Davídková, E. 1986. Nonparametric evaluation of graphical sensory profiles for the analysis of carbonated beverages. Die Nahrung, 30, 131–139.

Poulton, E. C. 1989. Bias in Quantifying Judgments. Lawrence Erlbaum, Hillsdale, NJ.

Richardson, L. F. and Ross, J. S. 1930. Loudness and telephone current. Journal of General Psychology, 3, 288–306.

Rosenthal, R. 1987. Judgment Studies: Design, Analysis and Meta-Analysis. University Press, Cambridge.

Shand, P. J., Hawrysh, Z. J., Hardin, R. T. and Jeremiah, L. E. 1985. Descriptive sensory analysis of beef steaks by category scaling, line scaling and magnitude estimation. Journal of Food Science, 50, 495–499.

Schutz, H. G. and Cardello, A. V. 2001.. A labeled affective magnitude (LAM) scale for assessing food liking/disliking. Journal of Sensory Studies, 16, 117–159.

Siegel, S. 1956. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill, New York.

Sriwatanakul, K., Kelvie, W., Lasagna, L., Calimlim, J. F., Wels, O. F. and Mehta, G. 1983. Studies with different types of visual analog scales for measurement of pain. Clinical Pharmacology and Therapeutics, 34, 234–239.

Stevens, J. C. and Marks, L. M. 1980. Cross-modality matching functions generated by magnitude estimation. Perception and Psychophysics, 27, 379–389.

Stevens, S. S. 1951. Mathematics, measurement and psychophysics. In: S. S. Stevens (ed.), Handbook of Experimental Psychology. Wiley, New York, pp. 1–49.

Stevens, S. S. 1956. The direct estimation of sensory magnitudes—loudness. American Journal of Psychology, 69, 1–25.

Stevens, S. S. 1957. On the psychophysical law. Psychological Review, 64, 153–181.

Stevens, S. S. 1969. On predicting exponents for cross-modality matches. Perception and Psychophysics, 6, 251–256.

Stevens, S. S. and Galanter, E. H. 1957. Ratio scales and category scales for a dozen perceptual continua. Journal of Experimental Psychology, 54, 377–411.

Stoer, N. L. and Lawless, H. T. 1993. Comparison of single product scaling and relative-to-reference scaling in sensory evaluation of dairy products. Journal of Sensory Studies, 8, 257–270.

Stone, H., Sidel, J., Oliver, S., Woolsey, A. and Singleton, R. C. 1974. Sensory Evaluation by quantitative descriptive analysis. Food Technology, 28, 24–29, 32, 34.

Teghtsoonian, M. 1980. Children's scales of length and loudness: A developmental application of cross-modal matching. Journal of Experimental Child Psychology, 30, 290–307.

Thurstone, L. L. 1927. A law of comparative judgment. Psychological Review, 34, 273–286.

Townsend, J. T. and Ashby, F. G. 1980. Measurement scales and statistics: The misconception misconceived. Psychological Bulletin, 96, 394–401.

Vickers, Z. M. 1983. Magnitude estimation vs. category scaling of the hedonic quality of food sounds. Journal of Food Science, 48, 1183–1186.

Villanueva, N. D. M. and Da Silva, M. A. A. P. 2009. Performance of the nine-point hedonic, hybrid and self-adjusting scales in the generation of internal preference maps. Food Quality and Preference, 20, 1–12.

Villanueva, N. D. M., Petenate, A. J., and Da Silva, M. A. A. P. 2005. Comparative performance of the hybrid hedonic scale as compared to the traditional hedonic, self-adjusting and ranking scales. Food Quality and Preference, 16, 691–703.

Ward, L. M. 1986. Mixed-modality psychophysical scaling: Double cross-modality matching for "difficult" continua. Perception and Psychophysics, 39, 407–417.

Weiss, D. J. 1972. Averaging: an empirical validity criterion for magnitude estimation. Perception and Psychophysics, 12, 385–388.

Winakor, G., Kim, C. J. and Wolins, L. 1980. Fabric hand: Tactile sensory assessment. Textile Research Journal, 50, 601–610.

Yamaguchi, S. 1967. The synergistic effect of monosodium glutamate and disodium 5′ inosinate. Journal of Food Science 32, 473–477.