Food Quality and Preference 22 (2011) 733–747

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Hedonic scaling: A review of methods and theory

Juyun Lim *

Department of Food Science and Technology, Oregon State University, Corvallis, OR 97331, United States

ARTICLE INFO

Article history:
Received 19 January 2011
Received in revised form 17 May 2011
Accepted 30 May 2011
Available online 6 June 2011

Keywords:
Hedonics
Scaling
9-Point hedonic scale
Labeled hedonic scale
Context effect

ABSTRACT

In recent years, interest in measuring hedonic responses has grown tremendously in both basic psychophysics and applied food and consumer research, resulting in the development of several new hedonic scaling methods. With these developments have come questions about theoretical and practical differences among the methods. The goal of this review is to compare and contrast these different approaches for the purpose of aiding researchers in selecting the most appropriate scaling method for their specific measurement needs. The review begins by addressing fundamental issues in scaling methodology, including the role of context effects, then moves on to describing and discussing the development of various types of hedonic scales, their specific properties, and their potential advantages and disadvantages.
Published by Elsevier Ltd.

Contents

* Tel.: +1 541 737 6507; fax: +1 541 737 1877.
E-mail address: juyun.lim@oregonstate.edu

0950-3293/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.foodqual.2011.05.008

1. Introduction

Sensations and hedonic experiences cannot be shared directly with others. For this reason, it was thought a century and a half ago that such experiences were inaccessible to direct measurement (see Boring, 1929; Savage, 1970). In search of valid methods of subjective measurement, Fechner (1860), who founded the science of psychophysics to study the relationship between physical stimuli and sensory responses, argued that sensory measurement could be best accomplished by measuring the subject's error in performing a discrimination task (e.g., responding to each test stimulus as "greater than" or "less than" the standard). It was not until years later that experimental psychologists, as well as consumer researchers, accepted the notion of using scaling methods, historically known as the "method of single stimuli" or "method of direct scaling", to measure sensory and hedonic responses. Many scaling methods have since been developed and have been used in a variety of situations to quantify sensation (e.g., Green, Shaffer, & Gilmore, 1993; Stevens & Galanter, 1957; Stone, Sidel, Oliver, Woolsey, & Singleton, 1974) and hedonic responses (e.g., Lim, Wood, & Green, 2009; Peryam & Pilgrim, 1957), as well as other perceptual and emotional dimensions, including satiety (e.g., Cardello, Schutz, Lesher, & Merrill, 2005), attitude (e.g., Likert, 1932), fear (e.g., Cox & Evans, 2008), and mood (e.g., Aitken, 1969; Zealley & Aitken, 1969).

Because the theoretical and practical differences among scaling methods are vast, there have been many studies, discussions and controversies in the scientific literature on the subject. The focus has, however, been primarily on measurement of sensation intensity. The development of hedonic scaling has lagged behind the development of intensity scaling, because measuring "secondary states", including evaluating the degree of liking/disliking of sensory stimuli, has generally been of less interest by psychophysicists (Prescott, 2009). In recent years, however, the importance of understanding human hedonics has become increasingly recognized in both consumer research and chemosensory neuroscience (de Araujo, Rolls, Kringelbach, McGlone, & Phillips, 2003; Rolls, Kringelbach, & de Araujo, 2003; Small et al., 2003; Winston, Gottfried, Kilner, & Dolan, 2005) and, as a consequence, interest in hedonic measurement has been greatly increased.

Although there have been several excellent reviews of psychophysical scaling, they have focused mainly on intensity scaling methods (e.g., Bartoshuk et al., 2002; Gescheider, 1988; Stevens, 1971). The current review focuses instead on scaling procedures that are designed to measure hedonic responses. Even so, it is necessary to begin by considering the fundamental issues and models that are common to both intensity and hedonic measurement, including the nature and function of different types of scales, and the role of context effects. After the background theory is presented, various hedonic scaling methods will be illustrated together with each scale's utility, properties, advantages, and disadvantages.

2. Fundamental issues in scaling methodology

2.1. Measurement types as a function of mathematical transformations

While each scaling method has its own unique features, the primary property of each method can be described by how numbers are conceptually utilized—for categorization, for ranking, for measuring degrees of difference, or for approximating magnitudes (Stevens, 1951). Listed in Table 1 are the basic operations, number usage, permissible statistics, and example hedonic scales for various scale types. In the nominal scale, which determines the identity of a measured property, the permissible transformation is the substitution of identifiers with numbers (e.g., stimuli A, B, and C, liked: "1" vs. stimuli D and E, disliked: "2"). The permissible transformation for data from an ordinal scale, which is designed to determine greater- or less-than relations between stimuli, consists of any increasing monotonic function. Under these conditions, the order, but not the degree of difference, will be preserved after such transformations (e.g., stimulus A, liked the most: "1", stimulus B, liked the second most: "2", and stimulus C, liked the least: "3"). An interval scale, which is designed to determine the equality of intervals (i.e., differences) between magnitudes, is invariant to any linear transformation in which the slope and intercept are free to vary. Thus, the interval scale has a variable unit size and an arbitrary zero. A ratio scale, which is designed to determine the equality of ratios among magnitudes, is invariant to linear transformations in which only the slope is free to vary and the intercept is zero. Thus, the ratio scale has a variable unit size and an absolute zero.

All of these scale types have been used in efforts to quantify hedonic responses. To fully understand both the theoretical and practical properties of such scales, the procedures used to construct them must be critically evaluated in relation to the type of data the scales are intended to yield.

2.2. Measurement theories and controversies

As noted above, the internal representation of sensory and hedonic experiences cannot be measured directly and so must be inferred from subjects' responses by means of descriptive or numerical data. The simplest conceptualization of sensory and hedonic measurements, therefore, involves two main stages of processing: sensory and cognitive. The general concept of a stimulus–response model (see Fig. 1), which was initially proposed to explain the discrepancy between the two psychophysical laws suggested by Fechner (1860) and Stevens (1957), has its origin in

Table 1
Scaling methods (adapted from Stevens, 1951).

Scale	Basic operation	Number usage	Permissible statistics	Example hedonic scales
Nominal	Categorization	Used as labels	Non-parametric: number of cases; Mode	1: good, 2: bad
Ordinal	Greater or less	Used to recognize the rank order	Non-parametric: median; percentiles	Rank rating
Interval	Differences	Used to represent degrees of differences	Parametric: mean; standard deviation	Category scale
Ratio	Ratios	Used to represent relative proportions	Parametric: log Mean; Standard deviation	LAM, LHS

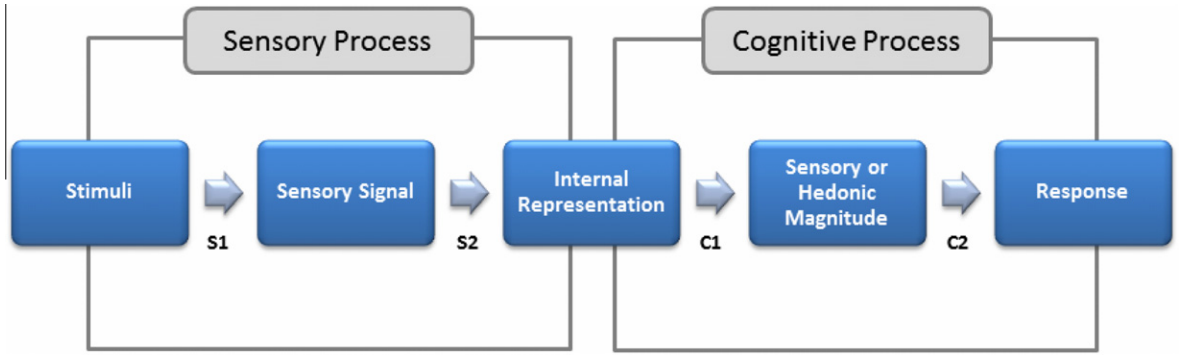

Fig. 1. A simple illustration of the stimulus–response model.

the work of Attneave (1962). Several other two-stage models, represented by two separate mathematical transformations – one for a sensory function and another for a response function – have also been proposed to explain the differences in psychophysical functions that were obtained by various experimental operations (e.g., category scaling vs. magnitude estimation) (Anderson, 1974; Birnbaum, 1982; Curtis, Attneave, & Harrington, 1968; Ekman, 1964; MacKay, 1963; Torgerson, 1961; Treisman & Williams, 1984; Ward, 1991). While the models of psychophysical judgment did not necessarily agree with one another in terms of mathematical functions (see Birnbaum, 1980), they all advanced understanding by pointing to the need to identify separate sources of the errors and biases that are inherent to sensory and/or hedonic measurement.

As illustrated in Fig. 1, the sensory input function comprises the first stage of processing, which includes sensory transduction of the stimulus via receptor processes (S1) and encoding to an internal representation (S2), including sensation quality, intensity and hedonic values. There is considerable evidence that the input function is strongly influenced by, among other things, the sensory perceptual context. The output stage includes evaluative and decisional processes involved in making a decision as to the appropriate response (C1) and producing that response (C2: e.g., assigning numbers, marking on a line). As shown in empirical studies (Beck & Shaw, 1961; Curtis et al., 1968; Torgerson, 1961), the output process can be modified by the response context, including the scaling method used (e.g., category scale vs. magnitude estimation) or by even the same scaling method with different instructions to subjects (e.g., magnitude estimation of stimulus intensities vs. of intensity differences). Thus, subjects map their internal representations onto the response continuum in different ways under different instructions, so that the same perceptual relation may be given either a "*difference* interpretation" or a "*ratio* interpretation" depending on whether *interval* or *ratio* judgments are requested (Beck & Shaw, 1961; Curtis et al., 1968; Torgerson, 1961). While there has been skepticism about whether human subjects can make ratio judgments, the accumulated evidence (which will be further discussed later) suggests that people are able to follow instructions to produce appropriate responses, although inevitably with some degree of error.

Indeed, any transformation, whether biological or cognitive, contains sources of error and bias. Given the framework suggested in Fig. 1, there are at least two points where some sorts of bias could arise: one for the sensory process and another for the response (cognitive) process. Historically, psychophysicists whose interests reside in elucidating sensory and perceptual mechanisms have emphasized the importance of avoiding response contexts that may distort the response transformation function. Ideally, for these purposes a scaling method should yield sensory or hedonic measurements that have a *linear* relationship to the internal representation of sensory event. Thus, they have tended to use scaling methods that contain a true zero and can, in theory, provide "absolute" data on perceptual magnitude. In the meantime, sensory scientists, whose objective is to compare the sensory and hedonic perception of products in the most objective and sensitive ways possible (e.g., descriptive analysis, consumer tests), have faced different challenges and found different solutions. Since the latter group is not interested in sensory processes per se, they have used scaling methods in more "relative" terms, focusing on the *differences* between sensory or hedonic magnitudes rather than on their absolute magnitudes.

These different viewpoints have created controversy over the validity and sensitivity of the different approaches (e.g., Mellers, 1983a,b; Zwislocki, 1983a; Zwislocki & Goodman, 1980). However, it is important to keep in mind that each scaling method is intimately tied to the issue of how we deal with two different contexts: the sensory context and response context. As Gescheider (1988) pointed out, the term "absolute", as used by Zwislocki and Goodman (1980) and Zwislocki (1983a) in describing magnitude estimation, does not mean that the scale cannot be biased during the response process. They used the term in accordance with Stevens' (1951) definition of scale types in terms of permissible mathematical transformations of the response function (Table 1). Arguing against Zwislocki, Mellers (1983a) emphasized that all psychophysical judgments, including absolute-magnitude estimates, are "relative" and occur in a (sensory) context. However, the disagreement between Zwislocki and Mellers is the result of two fundamentally different approaches to psychophysical scaling that are based on different goals, and are not amiss. Hence, it is important to understand how scaling methods differ from one another both theoretically and practically, where the potential biases may arise, and more importantly, what those biases means for sensory and hedonic measurements.

2.3. Context effects

Without doubt, sensory and hedonic perception is contextual. Hence, measurements of sensory or hedonic responses are inherently subject to context effects, which involve both sensory and response (cognitive) processes (Fig. 1). Context effects have been a central topic of interest in the field of sensory measurement, and a variety of effects on sensory and hedonic measurements have been studied (e.g., Algom & Marks, 1990; Diamond & Lawless, 2001; Helson, 1948; Parducci, 1974; Poulton, 1979; Schifferstein, 1995; Teghtsoonian, 1973). For example, numerous studies have reported changes in intensity and hedonic ratings and/or alterations in exponents of psychophysical function based on various experimental factors (e.g., number, spacing, and range of stimuli).

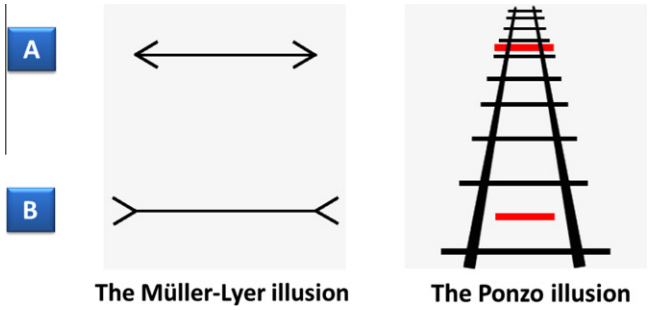

Fig. 2. Size illusions. In both illusions the horizontal lines, A and B, are of equal length.

These effects have sometimes been used as evidence against the validity of certain scaling methods, and have usually been interpreted as reflecting biases in response process. In other words, context effects have most often been considered as a distortion of the subject's response function rather than as an alteration in the internal representation of the stimulus itself. However, data suggest that context effects can be sensory in origin as well. That is, what may be interpreted to be a response bias may result from a change in the *internal representation* of the sensory input (See Fig. 2) rather than from a change in the output function.

2.3.1. Effects of sensory context

Compared to research on response biases, there is a relatively small number of studies that provide evidence that some common context effects (e.g., stimulus range and contrast effects) might be sensory in origin (Algom & Marks, 1990; Hulshoff Pol, Hijman, Baare, & van Ree, 1998; Marks, 1992; Marks & Warner, 1991; Ohzawa, Sclar, & Freeman, 1982; Parker, Murphy, & Schneider, 2002; Parker & Schneider, 1994; Schneider, Parker, & Moraglia, 1996). Schneider and Parker (1990) and Parker and Schneider (1994) presented subjects with two pairs of tones differing in intensity and had them select the pair with the larger loudness difference. The interesting finding in these two studies was that the subjects' judgments as to which pair had the larger loudness interval depended on the range of tones from which the pairs were selected. Because the subjects did not make numerical judgments, the context effect (i.e., range effect) could not be attributed to a numerical response bias. The authors instead speculated that sensory context can change the nature of the sensory representation through the operation of a gain-control mechanism. They suggested the notion of a nonlinear amplifier whose gain and degree of nonlinearity are adjusted under top-down control, so as to prevent distortion and increase discriminability.

Recently, we have also found evidence that participation in a taste detection task leads to higher ratings of the perceived intensity of suprathreshold taste stimuli even when the threshold measures were several days before the intensity ratings (Green & Lim, 2009). Because exposure to threshold-level taste stimulation specifically intensified suprathreshold taste perception, but not imagined taste sensations (e.g., the imagined bitterness of celery), the results were interpreted as a change in sensitivity or gain of the taste system rather than a response bias. It is unclear whether such a mechanism exists for hedonic perception. Nevertheless, it is important to keep in mind that context effects can be sensory or perceptual in origin, and not solely attributable to response bias. In fact, the best known and most classical visual illusions (Fig. 2) provide compelling evidence that the perception of stimuli results from an interaction between the properties of the sensory and perceptual systems and the context in which the stimuli occur.

Some of the context effects discussed below might be considered sensory in origin: that is, *perceptual biases* causing a change in internal representation. Those perceptual context effects can never be avoided and in fact their detection may attest to the sensitivity of the method being used rather than to a bias inherent to it. Some other context effects might be considered as response bias: that is, a shift or change *in response* to a constant percept (internal representation). While all scaling methods should be sensitive to context effects that affect sensory perception, some scales are more prone to response biases than others, which will be discussed below.

2.3.2. Contrast effects

It has been shown that the perceived intensity of a stimulus is rated as stronger in the context of weak stimuli and weaker in the context of strong stimuli (Lawless, 1983; Lawless, Horne, & Spiers, 2000; Mattes & Lawless, 1985; Rankin & Marks, 1991; Schifferstein & Frijters, 1992). This contrast effect, which can be considered to be either sensory or response-based in origin, occurs in hedonic perception as well. As Fechner (1898) suggested in his law of hedonic contrast, stimuli are liked less when they are sampled with better-liked stimuli (i.e., negative hedonic contrast) and are liked more when they are presented with less-liked stimuli (i.e., positive hedonic contrast) (Kamenetzky, 1959; Schifferstein, 1995; Zellner, Allen, Henley, & Parker, 2006).

Helson (1947, 1948) explained contrast effects in terms of adaptation level theory. Interestingly, the adaptation level effect does not refer to adaptation at the level of sensory processing (Stevens, 1975). Instead, the theory predicts that the average level of stimulation from the prior stimuli influences the following judgments at a behavioral level, such that exposure to strong stimuli results in subsequent underestimation of a test stimulus, while exposure to weak stimuli results in subsequent overestimation of the stimulus. Fundamental to the theory is that extreme stimuli change our "frame of reference" (i.e., range of stimulation) and thus the way it is mapped onto a response scale. In other words, because some scales (e.g., the 9-point hedonic scale) often do not specify a frame of reference (e.g., the context of foods in general vs. a specific food category), subjects have to gauge the context based on prior stimuli, which shifts the scale value for the next stimulus.

Zellner and colleagues (Zellner, Kern, & Parker, 2002; Zellner, Rohm, Bassetti, & Parker, 2003) recently demonstrated how subcategorization of stimuli reduces hedonic contrast by altering the frame of reference. In their experiments, subjects were instructed to consider the context of test stimuli either to be in the same category or in different categories. The contrast effect was attenuated for subjects who were instructed to view the context and test stimuli as being in different categories. For instance, dilute fruit juices that were followed by full-strength fruit juices were rated as less liked when subjects were told that all of the stimuli were fruit juices compared to when they were told that the dilute fruit juices were "commercial drinks" and the full-strength juices were "fruit juices". Thus, they have demonstrated that the size of the contrast effect can be manipulated by adjusting the frame of reference (i.e., by changing the assortment of stimuli subjects are asked to categorize). When the frame of reference is narrower (e.g., fruit juices instead of juice drinks), subjects tend to stretch the response scale and consequently the contrast effect seems greater. Such effects are prone to scales which do not provide explicit frames of reference. In another words, this artifact can be best avoided by using all-inclusive end anchors (see Section 3.3.2 below).

2.3.3. Range and frequency effects

Another prominent theory of relative judgments is that of Parducci (1965, 1974), which describes how category ratings are determined by the frequency distribution of the stimuli in a set.

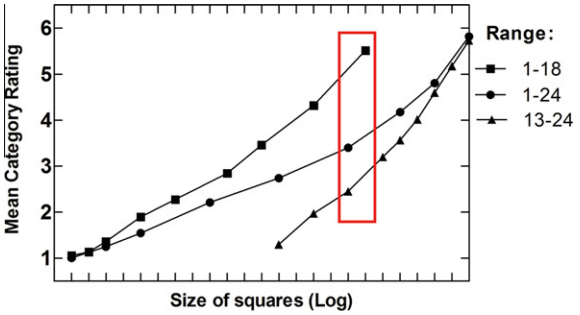

Fig. 3. Mean ratings of the sizes of squares in three stimulus sets, each having different ranges, rated on a 6-point category scale (produced from Parducci & Perrett, 1971, Table 1).

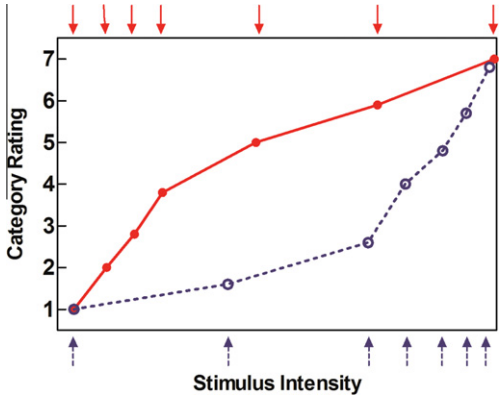

Fig. 4. Category ratings for two different stimulus spacing (regraphed from Galanter, 1966). Stimulus spacing indicated by arrows on the upper and the lower abscissa correspond to solid and dashed lines, respectively.

In his range-frequency theory, two principles, the range principle and the frequency principle, govern subjects' judgments on category scales. The range principle says that subjects tend to subdivide the available stimulus range into equal perceptual segments and assign sub-ranges to the available scale categories. For example, subjects who are given a 9-point category scale will divide the full range of stimuli into the nine approximately equal parts and use the ratings to identify those sub-ranges. Thus, a particular stimulus will be judged differently depending upon the range of stimuli in the set within which it is presented (Parducci, 1974; Parducci & Perrett, 1971). The data in Fig. 3 show a classical example of this effect (Parducci & Perrett, 1971), which is conceptually the same as the stimulus and response equalizing bias of Poulton (1979). In this example, three groups of subjects rated the size of squares on a 6-point category scale. Each group rated one of the three sets of squares composed of the same stimulus values, each with different stimulus ranges. The subjects used all or most of the available scale range, distributing the stimuli across the available responses as if the end of the category scale was stretchable to fit the end point of the stimulus range. Consequently, the psychophysical functions varied depending on the range of stimuli presented, such that the slope was steeper for the narrow range and flatter for the wide range. Teghtsoonian (1973), meanwhile, found the stimulus range to have little influence on magnitude estimation of apparent distance, apparent length, and loudness. This range effect is thus built more into the category scaling procedure, in which subjects are asked to place stimuli in a given number of categories.

The frequency principle holds that subjects have a tendency to use different parts of the scale equally often. In practice, this principle causes stimuli that are presented more frequently at a given sub-range to be spread out into neighboring categories. As seen in Fig. 4, the top and bottom curves show data generated from an experiment where the stimuli were spaced closer together towards either the lower end or the higher end of the stimulus range tested, respectively. As expected, the judgment curve was steepest at the top end of the stimulus range.

The results for hedonic measurement have been consistent with the relational nature of judgments in category scaling upon which the range-frequency model is based (Riskey, 1982; Riskey, Parducci, & Beauchamp, 1979). In the experiment by Riskey et al. (1979), subjects rated sweetness and pleasantness of soft drinks containing different concentrations of sucrose using 9-point category scales. The results clearly demonstrated the apparent frequency effect for both sweetness and liking: the same drinks were rated sweeter when the lower concentrations were presented more frequently and less sweet when the higher concentrations were presented more frequently, and the concentration producing peak pleasantness ratings (i.e., the breakpoint at the inverted U-shape psychohedonic function) was lower when the lower concentrations were presented more frequently.

In direct relation to the range-frequency theory, an important issue to consider is that the magnitude of the stimulus range and frequency effects are dramatically affected by both the number of categories a scale has and by the number of stimuli presented. Parducci (1982) and Parducci and Wedell (1986) demonstrated that the frequency effect is smaller when larger numbers of categories are employed. Conversely, the larger the number of stimuli presented the greater the frequency effects are (Parducci & Wedell, 1986).

Taken together, the range-frequency theory predicts that subjects compromise between these two tendencies so that their actual patterns of ratings reflect neither principle entirely, but fall between the predictions made by either principle alone. In addition, according to Parducci (1974) the range and frequency effects result from the subjects' tendency to distribute responses uniformly over the response continuum. Thus, such effects do not originate from sensory context, but from changes in the response function where the internal representation of sensations are converted to overt responses (Mellers & Birnbaum, 1982).

2.3.4. Other context effects

In addition to the range and frequency effects, other context effects related to category scaling appear to act by altering the response function. These include end effects, centering biases, and stimulus-spacing biases. It has been reported that when close-ended scales such as category and line scales are employed, the end points of the scales are used less frequently than other parts of the scale (i.e., end effects) (Anderson, 1974; Eriksen & Hake, 1957; Moskowitz, 1982; Schifferstein & Frijters, 1992; Stevens & Galanter, 1957; Yao et al., 2003; Yeh et al., 1998). Because subjects must consider the possibility that a better or worse (or stronger or weaker) stimulus may be presented later in the test, they are often reluctant to use the end points. Thus, a 9-point scale may effectively become a 7-point scale. O'Mahony (1982) explains this tendency in terms of psychological distances between categories: even though the intervals on category scales are intended to be equal, the psychological distance in traveling from the penultimate to the end category is greater than that for traveling between two categories in the center of the scale. This end effect may be especially important for results obtained with category hedonic scales, where only about half of the categories are allocated for each valence (e.g., four categories for each positive and negative valence in the 9-point hedonic scale).

The centering bias (Poulton, 1979) is a tendency for subjects to adjust the range of responses to the range of stimuli presented, causing the mean response to be centered in the middle of the response scale, regardless of its value. In the stimulus spacing bias, the subject responds as if the stimuli were subjectively equally spaced regardless of their actual perceptual spacing. Poulton (1979) points out that the centering bias and the stimulus-spacing bias form the basis for Parducci's range-frequency model (Parducci, 1965; Parducci & Perrett, 1971), which is an extension of Helson's (1948) original adaptation-level model. Even though those two biases are not commonly discussed, they may play a critical role in situations where there is a need for interpolating values (e.g., finding the optimum stimulus concentration) from a psychophysical or psychohedonic function. Further, such an effect may cause serious issues in comparing the magnitude of hedonic responses to stimuli across a group of subjects.

2.3.5. Dealing with context effect

Given the ubiquitous nature of context effects, it is necessary to be aware of their origin and causes so they may be taken into account or controlled when they influence the sensory input function, or may be minimized when they influence the response output function. The latter involves selecting a suitable scaling method that is less prone to context effects. There are a few practical approaches commonly used to handle context effects in general (for review, see Lawless & Heymann, 1998). The first is the use of different random or counterbalanced stimulus orders to minimize context effects which originate from the sensory input process. Immediate contrast effects between any two stimuli and simple order effects can be canceled out across subjects by using a sufficient number of orders, even though such a practice may not undo the broader effects of context for a given experiment, nor prevent scale-originated effects.

There are common approaches to reducing context effects related to the response function. The first is to calibrate subjects so that their frame of reference for the scale is internalized through instructions and a training process. While the "calibration" often refers to the intensive panel training associated with descriptive analysis with standard references, the central idea is to establish a constant frame of references for subjects so that ratings can be made in the same context across the subjects. As will be discussed more in the next section (see Section 3.3), some scales intend to provide a built-in frame of reference by using specific end-anchors. In cases where a scale itself does not offer a specific frame of reference, the rating scale is arbitrary and thus using a relative strategy is expected. For such cases, pre-exposure to the range of test stimuli has been recommended in order to establish a stable context during the scaling task (Diamond & Lawless, 2001). The second approach is stabilizing the experimental context across sessions containing stimuli that will be compared. It is often tempting to compare ratings given to a stimulus in different settings or from different experimental sessions. However, unless the context was the same in both experimental sessions, it is impossible to say whether differences between ratings arose from true sensation/hedonic differences or from contextual differences. Common practices include presenting warm-up or practice stimuli, or adding a "throw-away" stimulus. Lastly, most of the response-related context effects (e.g., range and frequency effects) are harder to deal with or even unavoidable (Poulton, 1979). In such cases, the best practice is to be aware of the potential response biases and to make proper inferences from the findings. For example, when subjects and consumers are given a category scale without training with the scale (or standard materials), they will use a relative strategy in a given experimental context. Accordingly, several scale-oriented response biases (e.g., end effects, centering bias) are expected. It is, therefore, inappropriate to make a statement about degree of liking/disliking from data obtained from such scale, while it is appropriate to make relative comparisons among stimuli.

3. Hedonic scaling

Over the last half century a number of scales have been developed and utilized to measure hedonic responses in both basic psychophysical and applied research. Some of the hedonic scaling methods will be discussed below together with each scale's utilities, properties, advantages, and disadvantages, as well as the history of its development.

3.1. The 9-point hedonic scale

3.1.1. History of development

Since its development, the 9-point hedonic scale (Peryam & Girardot, 1952; Peryam & Pilgrim, 1957) has been the most commonly used scale for testing consumer preference and acceptability of foods. Development of the scale, which began in 1947 at the Quartermaster Food and Container Institute for the Armed Forces, was motivated by the need for a rating scale that could overcome the limitations of the cumbersome method of paired comparisons (Peryam, 1950; Peryam & Pilgrim, 1957). The developers applied the graphic rating scale (Freyd, 1923; Guilford, 1936; Likert, 1932), which experimental psychologists had long used to measure various psychological phenomena, to measure the "hedonic value" (Peryam, 1950) of foods. In 1949, further preliminary work, in which scale lengths and wording were compared, was conducted and the present form of the scale (see Fig. 5) was selected based on its reliability and discriminability (Peryam & Pilgrim, 1957). The scale was introduced in 1952 (Peryam & Girardot, 1952) and it quickly became the method of choice by industry, government and academic researchers. However, the scale "was, perhaps, too immediately successful" and its success prevented further refinement of the scale, which the original developers had intended (Peryam & Pilgrim, 1957). In fact, in 1951 the Psychometric Laboratory at the University of Chicago was invited to evaluate the semantic meanings of hedonic phrases (Jones, Peryam, & Thurstone, 1955). The results suggested the psychological distances between the semantic labels on the 9-point hedonic scale were not equal (Jones et al., 1955; Jones & Thurstone, 1955), which was further confirmed later (Moskowitz, 1977, 1980; Moskowitz & Sidel, 1971). Unfortunately, those results were not utilized to refine the 9-point hedonic scale. Instead, the original form of the scale has been used since its development.

3.1.2. Properties

The 9-point hedonic scale is a balanced bipolar scale around neutral at the center with four positive and four negative categories on each side. The categories are labeled with phrases representing various degrees of affect and those labels are arranged successively to suggest a single continuum of likes and dislikes (Peryam & Pilgrim, 1957). The descriptors are intended to help not only subjects to respond accordingly but also to help experimenters interpret the mean value of responses in terms of degree of liking/disliking. One of the concerns about the scale during its development was whether its presentation format, i.e., long vs. short lines, vertical vs. horizontal orientation, or beginning with like vs. dislike, had effects on subjects' responses. It has been reported that such structural variations have no critical effect on the results (Peryam & Pilgrim, 1957). In terms of mathematical properties, the 9-point hedonic scale yields, in theory, ordinal data, since it is a category scale (i.e., ratings are limited to nine categories) with the labels that are spaced unequally in terms of psychological distances (Lim et al., 2009; Moskowitz, 1977, 1980; Peryam

(a)

	FOOD ITEM	LIKE				INDIFFERENT	DISLIKE			
Not Tried	Cream Gravy	Like Extremely	Like Very Much	Like Moderately	Like Slightly	Neither Like Nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely
Not Tried	Bread Pudding	Like Extremely	Like Very Much	Like Moderately	Like Slightly	Neither Like Nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely
Not Tried	Cheese	Like Extremely	Like Very Much	Like Moderately	Like Slightly	Neither Like Nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely
Not Tried	French Fried Onions	Like Extremely	Like Very Much	Like Moderately	Like Slightly	Neither Like Nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely
Not Tried	Lettuce Wedges	Like Extremely	Like Very Much	Like Moderately	Like Slightly	Neither Like Nor Dislike	Dislike Slightly	Dislike Moderately	Dislike Very Much	Dislike Extremely

(b) Overall, how much do you like or dislike this juice sample?

Sample 351

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

Fig. 5. Examples of the 9-point hedonic scale: (a) Questionnaire designed for studying soldier's preferences in the field (Peryam & Girardot, 1952); and (b) a sample ballot for a common consumer test used in a laboratory setting.

& Pilgrim, 1957). However, scale responses are, in practice, treated as points on a continuum instead of categorical and discrete data, so that the user can employ parametrical statistics such as analysis of variance, which are more sensitive than non-parametric counter-parts (Peryam & Pilgrim, 1957).

3.1.3. Advantages

The primary reason for the wide acceptance of the 9-point hedonic scale is that, compared to other scaling methods (e.g., magnitude estimation), its categorical nature and limited choices make it easy for both study participants and researchers to use. Its simplicity further makes the 9-point hedonic scale suitable for use by a wide range of populations without an extensive training. [Note: see Lawless & Heymann, 1998 for a review of hedonic measurements for children.] For researchers, data handling of the 9-point hedonic scale is also easier than other techniques which require measuring lines or recording magnitude estimates that may include fractions, although this practical matter is of diminishing importance given the development of computerized programs. More importantly, it has been shown that simple category scales are as sensitive as other scaling techniques (e.g., line marking and magnitude estimation) in terms of discrimination power (Lawless & Malone, 1986a,b). Therefore, when the primary concern of a study is measuring hedonic *differences* among foods, beverages, and consumer products and predicting their *acceptance*, the 9-point hedonic scale has proven itself to be a simple and effective measuring device.

3.1.4. Limitations

Despite its wide use in the field of sensory science, various limitations of the 9-point hedonic scale have been reported. First, as noted above, due to its inequality of scale intervals and the lack of a zero point (Moskowitz & Sidel, 1971; Peryam & Pilgrim, 1957), the scale can yield only ordinal- or, at best, interval data (i.e., ordered metric). Thus the scale cannot provide information about ratios of liking/disliking for stimuli (Moskowitz & Sidel, 1971; Schutz & Cardello, 2001) nor provide meaningful comparisons of hedonic perception between individuals and groups (Lim et al., 2009). Nevertheless, this does not pose problem for measuring relative (ordinal) preferences among stimuli, which was its intended purpose. Second, due to its limited number of response categories, the 9-point hedonic scale offers little freedom for subjects to express the full range of their hedonic experiences (Marchisano et al., 2003; Villanueva & Da Silva, 2009; Villegas-Ruiz, Angulo, & O'Mahony, 2008). Third, because of both its small number of available categories and the general tendency of subjects to avoid using extreme categories (Hollingworth, 1910; Moskowitz, 1982; O'Mahony, 1982), the scale is highly vulnerable to ceiling effects (Schutz & Cardello, 2001; Stevens & Galanter, 1957), one of the context effects that was described above (Section 2.3.4). The avoidance of the end categories effectively reduces the 9-point scale to a 7-point scale (Moskowitz, 1982; Moskowitz & Sidel, 1971) and limits its ability to discriminate among very well liked or very disliked stimuli (Lim & Fujimaru, 2010; Schutz & Cardello, 2001; Villanueva & Da Silva, 2009). Lastly, from a statistical standpoint, because the data it yields are categorical and discrete without a true zero point, the type of statistical analyses that can be applied with confidence is limited, i.e., nonparametric statistics. However, it is common practice for researchers to use more powerful parametric statistics, such as analysis of variance, to analyze data collected with the scale, although it is mathematically inappropriate to do so. In addition, as recognized in one of the original publications of the scale (Peryam & Pilgrim, 1957), some of the assumptions for parametric analyses (e.g., normality, homogeneity

of variance) are often violated (Gay & Mead, 1992; Giovanni & Pangborn, 1983; O'Mahony, 1982; Villanueva, Petenate, & Da Silva, 2000), particularly the data for extremely liked or disliked stimuli (Lim & Fujimaru, 2010; Lim et al., 2009). Accordingly, a large sample size, commonly over 75 responses per stimulus, is necessary to approximate normality in order to make valid statistical inferences.

3.2. Magnitude estimation

3.2.1. History of development

Stevens (1956, 1957) revolutionized the measurement of sensory magnitudes in the 1950s by promoting and developing ratio scaling methods. The most widely used of these methods is magnitude estimation (ME), which was originally called the method of absolute judgment. In its simplest form, subjects are asked to assign numbers to sensations that reflect the ratios of their perceived intensities. For example, a sensation twice as intense as another should be assigned a number twice as large. In the earliest study, Stevens instructed naïve subjects to assign numbers to the brightness of lights and to the loudness of sounds at different radiant and acoustic energy levels, respectively (Stevens, 1953). The results showed that: (1) remarkably similar functions emerged from the brightness measurement and the loudness measurement; (2) the functions approximated a power function of the stimulus energy [$R = k \cdot I^N$ or $\log R = \log k + N \log I$ (where R = the magnitude estimate of perceived intensity, k = a constant, I = radiant *or* acoustic energy level, N = the exponent)]; and (3) the data appeared to possess ratio properties. Stevens further conducted a variety of validation experiments that supported the idea that ME yields ratio level data (Stevens, 1955, 1956, 1957, 1974).

Magnitude estimation was applied to hedonic measurement first by Engen and McBurney (1964). These authors, who evaluated the pleasantness of a wide range of odors, using both ME and a 9-point category scale, found that the hedonic range of odors far exceeded the range of perceived odor intensities (i.e., 125/1 for hedonics vs. 2–3/1 for intensity). The early 1970s saw increasing use of ME for assessing likes and dislikes, first for model systems (Henion, 1971; Moskowitz, 1971), and later for actual foods (Moskowitz, Kluter, Westerling, & Jacobs, 1974; Moskowitz & Sidel, 1971). In fact, the latter studies compared performance of ME against the 9-point hedonic scale and showed that ME was as sensitive, if not more sensitive, than category scaling in terms of finding stimulus differences. In these early studies the subject assigned numbers on a unipolar scale, thus a magnitude estimate of 0 represented "no liking at all" or "unpleasant".

Bipolar hedonic magnitude estimation was first used by Moskowitz, Dravnieks, & Klarman (1976) in a study of the intensity and pleasantness of odors. In this method, positive and negative numbers are used to signify ratios or proportions of liking/disliking. Moskowitz went on to use this method in a study designed to optimize the acceptability of cola flavored beverages sweetened with artificial sweeteners (Moskowitz, Wolfe, & Beck, 1978). In the 1980s, ME became more popular as a method for measuring hedonic responses to foods, beverages and consumer products (e.g., Giovanni & Pangborn, 1983; McDaniel & Sawyer, 1981; Vickers, 1983; Warren, 1981). However, the method failed to overtake the 9-point hedonic scale, largely because of limitations such as the lack of semantic information and, more importantly, the difficult numerical nature of the task, particularly when applied in studies using untrained consumers.

3.2.2. Properties

Unlike any other psychophysical rating scales, the method of magnitude estimation does not depend on visual or semantic aids. Instead, the method asks subjects to assign numbers to sensory stimuli, without restriction, so that the ratios of the numerical assignments reflect ratios of sensory perceptions or of hedonic magnitudes. Just like other hedonic scales, the most commonly used form of hedonic ME (Moskowitz, 1982) is a bipolar scale which comprises positive numbers for likes, negative numbers for dislikes, and an intermediate value of 0, reflecting a neutral point. Numbers on opposite sides away from the center show increasing levels of likes or dislikes. Magnitude estimates on the same side of the scale can be easily interpreted in terms of direct ratio comparisons. For example, a +100 and a +25 mean that one is liked four times more than the other. Numbers on opposite sides cannot be as easily compared. However, for practical purposes, researchers often treat the positive and negative sides of the scale as equal and opposite, in which case a +100 and a −100 are equal and algebraically opposite (Moskowitz, 1977).

3.2.3. Validity of magnitude estimation

While ratio properties of ME are highly desirable, there has been substantial controversy concerning its validity (e.g., Anderson, 1982; Attneave, 1962; Birnbaum, 1980; MacKay, 1963; Treisman, 1964). At the heart of the controversy is the question of whether the psychophysical law achieved by ME reflects the relationship between stimulus intensity and sensory/hedonic magnitude or merely describes the relationship between stimulus intensity and the *judgment* of sensory/hedonic magnitude. In other words, researchers have questioned whether numerical judgments are directly proportional to sensory/hedonic magnitude [i.e., a potential nonlinearity of the response transformation function (C2, Fig. 1)]. While there is no objective way to prove linearity of the transformation function, work on sensation magnitude matching and additivity of measurements indicates that, at least for data averaged over several subjects, the response transformation function is linear. First, cross-modality and within-modality matches, in which subjects adjust the intensities of stimuli from different modalities or of qualitatively different stimuli (e.g., different frequencies of sound) to match their sensation magnitudes, have successfully predicted data from ME (Daning, 1983; Gescheider & Joelson, 1983; Hellman, 1976; Hellman & Zwislocki, 1964; Marks, 1966; Verrillo, Fraioli, & Smith, 1969). Moreover, several experiments have supported the hypothesis that magnitude estimates are additive measures of sensation magnitude (Cain, 1976; Dawson, 1971; Hellman & Zwislocki, 1963, 1964; Marks, 1978, 1979; Marks & Bartoshuk, 1979; Zwislocki, 1983b). In these cases, the average magnitude estimate of two stimuli presented together was found to be equal to the sum of the average magnitude estimates of the stimuli presented alone. In addition, the fact that the quantitative relationships among perceived intensities and hedonic magnitudes of the same stimuli measured by ME were virtually identical to those obtained by category-ratio scaling (see below) (Green et al., 1993; Lim et al., 2009) further adds credence to both scaling techniques. In fact, the minimal requirement for the validity of a psychophysical scale is that two stimuli that have the same scale values should be judged to be subjectively equal; in these studies the ratios of subjective magnitude were also equal. While validity of ME remains to be controversial, above listed evidence is quite notable.

A very much related yet more practical controversy regarding ME is whether subjects can *accurately* estimate sensation ratios. Specifically, the tendency for subjects to use round numbers, such as 5, 10, 20, 100, etc., has been used as evidence that ME is not valid. While the "round number" bias (Giovanni & Pangborn, 1983; O'Mahony & Heintz, 1981; Stevens, 1975) certainly occurs, especially with subjects who are less trained on the task, it is not proof that ratios cannot be estimated. Such a bias can be viewed instead as reducing the *resolution* of ratio ratings. Accordingly, various

experimental procedures (Moskowitz, 1977; Stevens, 1975) have been suggested to reduce the effects of this bias.

3.2.4. Advantages

Magnitude estimation provides some key strength over category or line-marking scales. The primary advantage is that inferences can be made about the differences in liking/disliking among stimuli or sensations in terms of ratios. That is, ME better illustrates the relationship between changes in physical intensity and overall liking/disliking than does the traditional fixed point category scale. This can be a very useful tool in both basic research as well as sensory evaluation of foods. For example, hedonic magnitudes can be measured for different groups of individuals (e.g., with or without taste or olfactory disorders) and the comparisons can be used to understand normal vs. pathological sensory systems. Likewise, product comparisons can be made while other market variables are considered (e.g., using a premium ingredient costs 10% more than a regular ingredient, but will result in a 20% increase in consumer liking). In addition, ME reveals differences between stimuli just as well, and in some instances better than category scaling (Lawless & Malone, 1986b; McDaniel & Sawyer, 1981; Moskowitz & Sidel, 1971; Pearce, Korth, & Warren, 1986; Shand, Hawrysh, Hardin, & Jeremiah, 1985; Vickers, 1983), especially when large number of stimuli or a few very well liked stimuli are being tested in one experiment.

3.2.5. Limitations

Of course ME also has several important limitations. First, since all judgments are made relative to one another and subjects can choose their own numbers, there is no provision for anchoring the judgments of individual subjects to a common ruler, i.e., there is no certainty that a rating of '9' means the same to all subjects (but see Zwislocki & Goodman, 1980). Therefore, direct comparisons of rated values between subjects are meaningless. Second, the absence of semantic information (e.g., "like very much") prevents researchers from translating ratio differences into useful comparisons of differences in product perception of liking. Most importantly, as mentioned above the numerical nature of the task, which involves using numbers to estimate ratios, can be very difficult for naïve subjects. This means that the quality of the data obtained with ME often depends on the level of experience or training with the method that subjects have. For example, some loss of sensitivity was found in using ME with untrained heterogeneous samples of consumers, while this was not the case with untrained college students (Lawless & Malone, 1986a,b). Finally, the complexity of analyzing the data can be another hurdle: it requires normalization and standardization even before starting statistical analyses (see Moskowitz, 1977). Magnitude estimation is therefore more cumbersome to use than other scaling methods and requires detailed instructions and practice that is not possible to provide in some experimental situations, especially those involving consumers. For these reasons, ME has never been widely adopted in applied sensory research.

3.3. Category-ratio scales

3.3.1. The origin of category-ratio scales

While the debate about the validity and the superiority of category vs. ratio scaling continued, Borg (1982) proposed a new type of rating scale, called a 'category-ratio' scale, which as its name implies adopted positive features from both scaling methods. Simply put, a category-ratio scale is a line scale that has verbal descriptors of magnitude placed at selected positions along the line in such a way that it yields ratio-level data. Borg (1982) reasoned that "...category methods... are very popular for practical use..." although "...they offer no possibilities of direct ratio comparisons of perceptual intensities." He went on to explain that "...the ratio methods seem to give a better representation of the relative perceptual variation than other scaling methods, where direct intensity estimates can be obtained." However, "...the ratio methods only give relative intensities and no subjective 'levels' for immediate inter-individual or inter-modal comparisons..." "With category judgments, on the other hand, the intensities may be judged and evaluated in a more 'absolute' sense, i.e., direct 'level estimates' may be made from the intensities, whether they are 'strong' or 'weak', according to the life-long experience of the individuals or fundamental psychophysiological responses" (p. 25–26).

Several important observations and assumptions underlie the development of the category-ratio scale: The first is the acceptance of the view (Stevens, 1957) that the method of magnitude estimation yields ratio-level data. Borg (1982) acknowledged that magnitude estimates of sucrose and citric acid made by two patients who underwent inner ear surgery were highly correlated with their neural responses (Borg, Diamant, Strom, & Zotterman, 1967) and used this evidence as a validation of the ratio scaling method. The second is the assumption that the perceptual range is the same for all individuals (Borg, 1961) and all modalities (Borg, 1994; Teghtsoonian, 1971, 1973), although the physical range of the stimulus may vary considerably (i.e., range theory; Borg, 1961, 1970, 1971). For example, he believed that individuals experience the same degree of subjective exertion when they perform dynamic work at their respective maxima, although different people may need different physical workloads to achieve this maximum. If this is so, then all individuals are "calibrated" to the same maximal exertion. The third consideration is the empirical evidence (Borg & Hosman, 1970; Borg & Lindblad, 1976; Hosman & Borg, 1970) that adjectives and adverbs, which possess psychological magnitudes, can be used to define the 'level' of certain perceptual intensities, and that semantic meanings can be experimentally determined on a 'ratio level'.

Based on this foundation, Borg (1982) derived a category-ratio scaling by using the linear relation between (1) a commonly used category scale for *ratings of perceived exertion* (i.e., RPE scale) vs. the physical work load, and (2) magnitude estimates of perceived exertion vs. the work load. Unfortunately, early tests of the scale with sensations other than physical exertion (e.g., taste) produced somewhat disappointing results, which led to modification in the locations of the semantic descriptors (Borg, 1982). The modified version of the scale comprises nine descriptors, from "no sensation" to "maximum sensation", which are roughly linearly spaced along a logarithmic numerical scale. Using the scale, Borg and his colleagues compared psychophysical functions for exertion, taste and loudness with functions obtained with magnitude estimation. The results showed that agreement between the methods was great for perceived exertion but not as good for the other modalities, even after the adjustments of label locations had been made in various ways (Borg, 1982, 1990; Borg & Borg, 1987; Borg, Ljunggren, & Marks, 1985; Marks, Borg, & Ljunggren, 1983).

The questionable performance of the category-ratio scales for modalities other than exertion motivated Green et al. (1993) to develop another category-ratio scale to measure the perceived intensities of oral sensations. Instead of relying on the assumption that the perceptual range is the same for all sensory modalities, Green and his colleagues constructed a scale using ratings of the perceived magnitudes of semantic descriptors obtained within the context of the modalities of interest, which were somesthesis and gustation. The scale was thus constructed by asking subjects to estimate intensity magnitudes of verbal descriptors (e.g., 'weak', 'strong') that were presented along with examples of a variety of common oral sensations including oral pain. The resulting scale, which was called the oral labeled magnitude scale (LMS), is bounded by 'no sensation' at the bottom and 'strongest imaginable

oral sensation of any kind' at the top, with five more descriptors spaced in quasi-logarithmic manner. In the same study, the LMS was also evaluated to determine whether it could produce intensity data comparable to the method of magnitude estimation for a variety of oral sensations (i.e., taste, chemesthesis, and temperature). The psychophysical functions generated by the two methods were statistically indistinguishable, indicating that the LMS yielded ratio-level data comparable to that produced by ME on the intensity of diverse oral sensations when the sensations were experienced and evaluated within a common perceptual context. Green et al. (1996) subsequently established that the LMS also produced data equivalent to magnitude estimation for odors as well as for tastes, as long as subjects made their ratings in the context of all possible tastes and smells, including painful 'tastes' (e.g., chili pepper) and 'odors' (e.g., ammonia). This finding led to the conclusion that the LMS could be used for any perceptual continua on which the strongest imaginable sensations were painful.

In later years, Bartoshuk and colleagues (Bartoshuk, 2000; Bartoshuk et al., 2002; Bartoshuk, Duffy, Green et al., 2004) argued that the top of the LMS (i.e., strongest imaginable oral sensation of any kind) differs across individuals due to the differences in tongue anatomy, and thus that anchoring the top of the scale with a domain which is being tested (i.e., taste in this case) could produce a ceiling effect for those who have high sensitivities in that domain. They further argued that, unlike Borg and Teghtsoonian assumed in earlier years, intensity maxima are not the same across modalities [e.g., maximum sweetness vs. pain, as noted by Green et al. (1993)]. Based on this argument and empirical data (Bartoshuk et al., 2002), the top anchor of the LMS was replaced with 'strongest imaginable sensation of any kind', and the scale was referred as the general version of the LMS, or gLMS.

3.3.2. Extension of category-ratio scales to hedonic measurements

In recognition of its positive features (see Section 3.3.3.), many researchers have recently adapted category-ratio scaling for measurement of hedonic responses. The first was the labeled affective magnitude (LAM) scale for assessing food liking/disliking (Schutz & Cardello, 2001), which was derived specifically for hedonic experiences associated with foods. Following a generally similar psychophysical procedure used to create the LMS, the authors asked a group of subjects to rate 44 semantic labels for their affective meaning in the context of foods using modulus-free magnitude estimation. Based on the results of multiple studies in which they assessed the effects of alternative semantic and numeric labels on the sensitivity and reliability of a potential scale, the LAM scale was derived with verbal labels that are consistent with the 9-point hedonic scale, with two additional anchors: 'greatest imaginable like' and 'greatest imaginable dislike' (see Fig. 6). About the same time, the need for a scaling method which could quantify individual and group differences in hedonic responses led some other researchers to adapt the gLMS to a bipolar hedonic scale using the same adjectives and spacing (Bartoshuk, Duffy, Chapo et al., 2004). Based on the assumption that hedonic magnitude and perceived intensity have a similar scalar structure, the bipolar gLMS was constructed with 'neutral' at its center-point and with positive and negative ratings on each side (see Fig. 7). A third hedonic category-ratio scale is the 'Oral Pleasantness and Unpleasantness Scale' (OPUS) (Guest, Essick, Patel, Prajapati, & McGlone, 2007). These authors' rationale for developing the OPUS was that some of the adjectives used to describe perceived intensity on the gLMS are inappropriate for pleasantness/unpleasantness ratings, and that spacing among descriptors might be different for specific sensations, such as oral pleasantness, wetness, roughness, etc. Thus, similar to the LAM scale, they used basically the same approach that was used to develop the LMS, but within a narrow semantic context of oral sensations, with painful sensations purposely avoided. Interestingly, this strategy resulted in a semantic structure similar to the LAM scale (see Fig. 7).

Fig. 6. The labeled affective magnitude (LAM) scale (Schutz & Cardello, 2001).

More recently, Lim et al. (2009) developed yet another hedonic category-ratio scale in recognition of some potential limitations of the existing scales. First, the frames of reference for the LAM scale and OPUS were limited to either foods or oral sensations. While developing a scale within a corresponding context (e.g., foods) may provide an appropriate frame of reference to make comparisons for items of these kinds, it is unclear if the scale provides a valid context for other hedonic experiences (e.g., non-food items) or provides valid individual and group differences in hedonic perception of foods (For the arguments made for valid vs. invalid comparisons across individuals and groups, see Bartoshuk et al., 2002). Secondly, some aspects of the psychophysical procedures employed to derive the existing hedonic scales, such as the amount of experience subjects had with magnitude estimation, differed from those used to develop the LMS, raising questions about their validity. Finally, because the gLMS was intended to measure sensory intensities, its descriptors, with the exception of 'moderate', do not translate very well to hedonic measurements (e.g., barely detectable). In addition, when the spacing among descriptors of the bipolar gLMS was compared with those of the other two hedonic category-ratio scales, 'moderate' on the bipolar gLMS was located much closer to neutral than was 'moderately' on the other two scales (see Fig. 7), which raised questions about potential differences in semantic structures underlying hedonic and intensity continua.

The labeled hedonic scale (LHS) (Fig. 8) was developed using a procedure that adhered closely to the procedure used to develop the LMS, and once developed it was directly compared against magnitude estimation and the 9-point hedonic scale (Lim et al., 2009). The results showed that the LHS yielded data that were almost identical to those obtained using magnitude estimation, supporting the validity of placement of the semantic descriptors and the assumption of ratio-level data. In addition, compared to the 9-point hedonic scale the LHS afforded slightly better discrimination among stimuli and much greater resistance to ceiling effects while producing more normally distributed data.

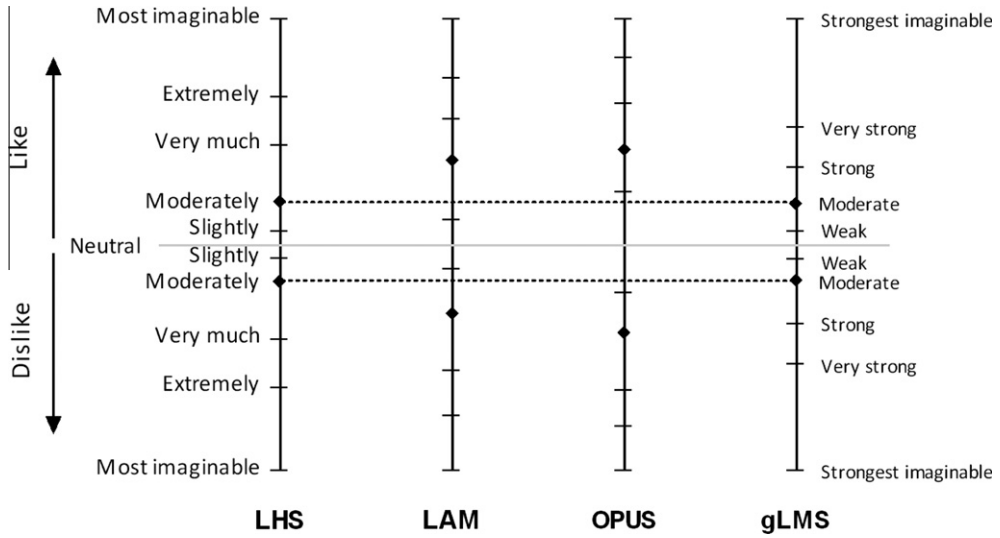

Fig. 7. Shown for comparison are the locations of semantic descriptors on the LHS, the LAM, the OPUS, and the bipolar gLMS (Lim et al., 2009). Filled diamonds indicate the location of 'moderately' on each scale, which is the only semantic descriptor other than neutral that is common to all of the scales. Horizontal dotted lines intersect the other scales at the locations of "like moderately" and "dislike moderately" on the LHS. Tick marks indicate the locations of the four other positive and negative descriptors that are "semantically equivalent" on the LHS, the LAM, and the OPUS. The remaining four descriptors of the gLMS, which have no direct counterparts on the other three scales, are shown on the right. The numerical values for the semantic labels can be found in the original papers for each scaling method (the LHS: Lim et al., 2009; the LAM: Schutz & Cardello, 2001; the OPUS: Guest et al., 2007; the gLMS: Green et al., 1993).

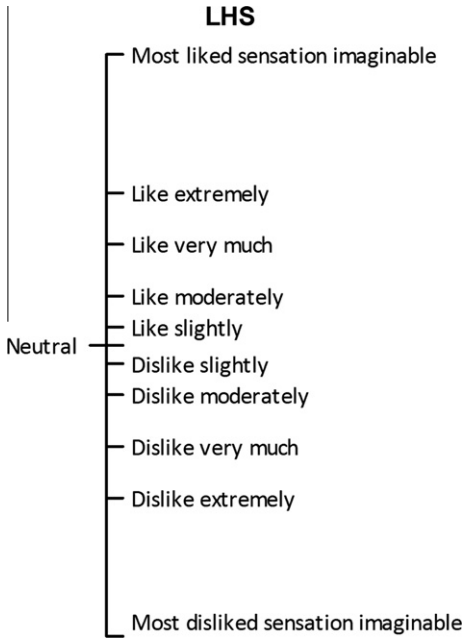

Fig. 8. The labeled hedonic scale (LHS) (Lim et al., 2009).

3.3.3. Properties and advantages

There are several distinctive features and properties of category-ratio scaling. First, because they were derived using ratio scaling, they can be assumed to yield ratio-level data equivalent to magnitude estimation, although only the LHS has been validated against magnitude estimation. This property is particularly valuable to illustrate the relation between hedonic magnitude or acceptability and underlying quantitative dimension of the stimulus (e.g., concentration of sweetener), and to make ratio statements about differences in liking (e.g., stimulus A is liked twice as much as stimulus B). Second, because the positions of their semantic labels have been empirically determined, they provide meaningful semantic information about subjective experience in addition to quantitative data. Interestingly, this property has not been considered as important as it should be: semantic information effectively translates numerical data into meaningful statements about the intensity or hedonic value of stimuli. Third, because they are continuous line scales, subjects can express subtle differences in preference among stimuli rather than being confined to categorical judgments. The fact that they have high-end anchors also increases the sensitivity of the scales for discriminating stimuli, especially very intense or highly liked or highly disliked stimuli (El Dine & Olabi, 2009; Greene, Bratka, Drake, & Sanders, 2006; Lim et al., 2009; Schutz & Cardello, 2001). Fourth, the hedonic gLMS and the LHS, because they are bounded by all-inclusive end anchors (e.g., 'Most liked or disliked sensation imaginable'), enable comparison of individual and group differences within the context of the full range of perceptual experiences (Bartoshuk et al., 2002). Fifth, unlike the intensity ratings obtained by magnitude estimation and the gLMS that are typically distributed log-normally across subjects (Green et al., 1993), it has been shown that the hedonic ratings obtained by category-ratio scaling are normally distributed (Lim & Fujimaru, 2010; Lim et al., 2009). Therefore, parametric analyses (e.g., ANOVA) can be readily applied without any data transformation, i.e., there is no need to log-transform the data obtained from the LHS. Finally, category-ratio scales have been shown to be as easy to use for subjects as the 9-point hedonic scale (Lim & Fujimaru, 2010; Schutz & Cardello, 2001), although they might require instructions and practice ratings to obtain the highest quality data (see below).

3.3.4. Limitations

While hedonic category-ratio scales can be an advantageous alternative for hedonic measurement of taste, flavor, foods, and potentially any other hedonic experiences, some concerns about category-ratio scales have also been raised. The most fundamental question for any category-ratio scale, particularly in the field of sensory evaluation of foods and consumer products, has been the possibility that the all-inclusive end anchors (i.e., most imaginable sensory experience of any kind) result in compression of ratings toward the center of the scale (i.e., neutral) (Cardello, Lawless, & Schutz, 2008), thus reducing the discrimination power of the scale.

The assumption underlying this concern is that the maximum liking or disliking of foods is far less than the most liked/disliked sensation of any kind. While this assumption might be true for some people, the data so far collected (Lim & Fujimaru, 2010; Lim et al., 2009) have shown that hedonic ratings for some extremely liked or disliked samples in fact come close to the end anchors, at least for a subgroup of subjects. More importantly, the results from previous studies showed that discrimination performance on hedonic category-ratio scales was no less than that of the 9-point hedonic scale (El Dine & Olabi, 2009; Greene et al., 2006; Lawless, Popper, & Kroll, 2010; Lim et al., 2009; Schutz & Cardello, 2001) even when a small number of stimuli, covering a relatively narrow hedonic range, were tested (Lim & Fujimaru, 2010). While response compression is generally considered an undesirable trait of any scale, it is important to emphasize that such a phenomenon itself does not necessarily mean that the sensitivity of the scale is poor. Data have suggested that although a wider frame of reference may reduce the range of the scale being used, resolution of the scale remains relatively unchanged because the variances decrease as mean ratings decrease (Cardello et al., 2008; Lim & Fujimaru, 2010).

What may be more problematic is the potential misuse of the scale by naïve subjects, and by subjects who have extensive experience with other scales, such as the 9-point hedonic scale. In a recent study, Cardello et al. (2008) reported that "a large number of panelists (50 of 100 panelists at one study site and 65 of 100 panelists at another site) used the LAM scale in a categorical manner, making ratings on the tick marks corresponding to the verbal labels" (p. 476). A similar improper use of the LHS was also seen in a large consumer test for subjects who had previous experience with the 9-point hedonic scale, and had not received detailed instructions emphasizing the use of the scale. Nevertheless, the categorical rating behavior was not evident among those who received the proper instructions (Lim & Fujimaru, 2010). Even though instructions about scale usage are often given a low priority, especially in large consumer tests, subjects' full understanding of the nature and use of a scale is of critical importance for obtaining valid data. In order to take full advantage of any scale, understanding the theory and properties of the scale as an experimenter and providing proper instructions to study subjects prior to testing is crucial.

3.4. Relative hedonic scaling

In a search of a superior alternative to the 9-point hedonic scale, some researchers found solutions in a completely different and fundamentally opposite approach. Instead of measuring the degree of hedonic reaction to stimuli in a broad context, and thus in a more absolute manner (i.e., like "very much" within a certain frame of reference), their approach was to measure the degree of hedonic relativity among stimuli. This approach assumes that when a subject assesses multiple stimuli, the rating of a stimulus is only relative to the ratings of the other stimuli (Cordonnier & Delwiche, 2008; Koo, Kim, & O'Mahony, 2002; Mellers, 1983a). Rank-rating (Kim & O'Mahony, 1998; O'Mahony, Park, Park, & Kim, 2004), also known as positional relative rating (PRR) (Cordonnier & Delwiche, 2008) is such a procedure. In this method, a category scale (e.g., 9-point or 21-point) is presented on a cardboard strip placed in front of the subject, who then rates the stimuli by placing each one in front of the appropriate number on the strip. As the subject proceeds through the test, they are allowed to retaste the stimuli as often as they wish and to alter the locations of stimuli that have been already positioned. While the initial experiment showed that the rank-rating procedure provided less discrimination errors in intensity measurements (e.g., saltiness of NaCl) compared to a 9-point category scale (Kim & O'Mahony, 1998), application of rank-rating in hedonic measurement did not show a significant advantage over the 9-point scale (Cordonnier & Delwiche, 2008; O'Mahony, Park, Park, & Kim, 2004). In addition, because rank-rating requires multiple tasting and retasting of each stimulus, this procedure may not be equally suitable for certain testing situation.

Another relative scaling method is the self-adjusting scale (Gay & Mead, 1992; Mead & Gay, 1995; Villanueva, Petenate, & Da Silva, 2005; Villanueva et al., 2000), in which subjects are required to place the most liked stimulus at the right end of the scale and the least liked stimulus at the left end of the scale, and then partition all of the others at appropriate intermediate locations. The advantages claimed for this method are that it demands the least training time for subjects, and that it eliminates differences among subjects in their usage of scale ranges, as everyone has to use the whole range of the scale. However, studies have suggested that the data generated from the self-adjusting scale show serious distortions from normality and that the scale is less efficient than the 9-point hedonic scale with the respect to discrimination power (Villanueva et al., 2000, 2005).

In recognition of the above mentioned limitations observed in both the 9-point hedonic scale and the self-adjusting scale, Villanueva and colleagues proposed another relative scale, called the hybrid hedonic scale (Villanueva et al., 2005). The hybrid hedonic scale is a linear scale resulting from the combination of the structured and unstructured scales. The scale has equidistant points and three verbal affective labels in the middle (i.e., neither liked nor disliked) and both ends (i.e., "disliked extremely" and "liked extremely") of the scale. The claimed advantages of this scale over the 9-point hedonic scale are: (1) because it is not restricted to a limited number of categories, the scale offers better discrimination power; (2) that it reduces the psychological error of habituation, and (3) because the scale generates continuous data, it allows for the use of parametric analyses (Villanueva & Da Silva, 2009; Villanueva et al., 2005). Such claims, however, have recently been challenged. Lawless (2010) argued that "the details of this paper do not justify any strong endorsement of the hybrid scale, nor any condemnation of the traditional 9-point hedonic scale". Thus, it is not yet clear whether the hybrid scale offers any real advantages over other hedonic scaling methods, including the 9-point hedonic scale.

3.5. Indirect hedonic scaling

The scaling methods discussed above (i.e., the 9-point hedonic scale, magnitude estimation, category-ratio scales, and relative scales) produce numbers that have face values (e.g., "like very much" on the 9-point hedonic scale = 8 on a 1 to 9 scale; "like very much" on the LHS = 44.43 on a −100 to 100 scale) which represent degrees of liking and disliking, and thus are considered direct scaling methods. Another approach to scaling is to use the variance created from relative judgments as units of measurement. This procedure is a very different approach in the sense that it does not derive scale values from a response "scale" per se, and accordingly cannot produce data in terms of *mean* response scale values. Instead, statistically-based variances from choice-based tasks (e.g., choose the one you prefer) are derived and used to construct a measurement scale. This form of "indirect scaling" (also called as Fechnerian methods, Baird & Noma, 1978; Jones, 1974) produce scale values which describe differences between stimuli in terms of how many *standard deviations* separate them. Such variability-based procedures are based on Thurstonian modeling (Thurstone, 1927a,b), which is well established in intensity measurement (see O'Mahony, Masuoka, & Ishii, 1994).

Recently, a variability-based procedure of hedonic measurement, called best-worst scaling (Jaeger, Jorgensen, Aaslyng, & Bredie, 2008), has been introduced to the field of sensory

evaluation. Originally used in studies of preferences for complex attitudinal dimensions (e.g., Finn & Louviere, 1992; Flynn, Louviere, Peters, & Coast, 2007; Marley & Louviere, 2005), the best-worst approach extends the method of paired preference tests by asking subjects to choose the best and the worst stimuli from a set of three or more stimuli. Simple difference scores are usually calculated based on the number of times a stimulus is called best (+1) vs. worst (−1), and these scores are assumed to have interval properties (Marley & Louviere, 2005). However, it has been theorized (Finn & Louviere, 1992; Marley & Louviere, 2005) that best-worst scaling can yield ratio-level data if a multinomial logistic regression is performed on the data. Just like any other forced-choice task, the best-worst scaling procedure has been shown to provide better discrimination power compared to direct scaling methods (Hein, Jaeger, Carr, & Delahunty, 2008; Jaeger et al., 2008), and to be easy to use by consumers because it simply requires them to choose the best-liked and worst-liked products in a series (Jaeger & Cardello, 2009). However, indirect measures of differences, including best-worst scaling, also have important limitations (see Moskowitz, 2005). First, the procedures only provide the relative degree of preference among a set of stimuli, not the degree of liking and disliking of each stimulus. For example, although the data may show a clear preference for one of the test stimuli over the others, there is no way to know whether that stimulus is "liked" or not, i.e., it yields no hedonic value. Second, the procedure can be very labor intensive. Each scale value has to be derived from multiple observations of choice experiments (e.g., ranking tasks), which means the procedure requires many more tasting trials than a direct scaling procedure (e.g., 21 tastings in a block design for seven stimuli), rendering it difficult to perform with foods and beverages (Jaeger & Cardello, 2009).

4. Conclusion

During the past decade, interest in measuring hedonic responses has grown tremendously in both basic psychophysics and applied food and consumer research. In the field of chemosensory science, studying individual and group differences in hedonic responses to chemical stimuli has become fundamental to reaching a better understanding of the role of sensory, perceptual, cognitive and genetic factors in food preference and selection. At the same time, in applied product research, discovering underlying consumer segmentation, (instead of just finding out which products are liked more than others) has become more essential than ever before as the consumer marketplace has become more crowded and competitive. These needs coupled with the recognition of the positive features of category-ratio scaling motivated the development of various hedonic category-ratio scales. Understandably however, there has been confusion about the theory and value of different scales among sensory scientists and professionals.

By describing the properties, advantages, and limitations of various scales I do not mean to suggest that one scale is necessarily a superior or inferior measuring instrument than others. There is no golden method which provides everything in a single click, nor are there methods which are not useful at all. Instead, the goal of this review is to aid researchers in the challenging task of identifying the most appropriate, sensitive and valid scale for the type of hedonic data they seek to collect in specific experimental contexts, and optimizing the quality of the data by using the appropriate procedures and instructions.

References

Aitken, R. C. (1969). Measurement of feelings using visual analogue scales. *Proceedings of the Royal Society for Medicine, 62*(10), 989–993.

Algom, D., & Marks, L. E. (1990). Range and regression, loudness scales, and loudness processing: toward a context-bound psychophysics. *Journal of Experimental Psychology: Human Perception and Performance, 16*(4), 706–727.

Anderson, N. H. (1974). Algebraic models in perception. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception. II. Psychophysical judgment and measurement* (pp. 215–298). New York, NY: Academic Press.

Anderson, N. H. (1982). Cognitive algebra and social psychophysics. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 123–148). Hillsdale, NJ: Lawrence Erlbaum associates.

Attneave, F. (1962). Perception and related areas. In S. Koch (Ed.). *Psychology: A study of a science* (Vol. 4). New York: McGraw-Hill.

Baird, J. C., & Noma, E. (1978). *Fundamentals of scaling and psychophysics*. New York: Wiley.

Bartoshuk, L. M. (2000). Comparing sensory experiences across individuals: Recent psychophysical advances illuminate genetic variation in taste perception. *Chemical Senses, 25*, 447–460.

Bartoshuk, L. M., Duffy, V. B., Chapo, A. K., Fast, K., Yiee, J. H., Hoffman, H. J., et al. (2004). From psychophysics to the clinic: Missteps and advances. *Food Quality and Preference, 15*, 617–632.

Bartoshuk, L. M., Duffy, V. B., Fast, K., Green, B. G., Prutkin, J., & Snyder, D. J. (2002). 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., et al. (2004). Valid across-group comparisons with labeled scales: the gLMS versus magnitude matching. *Physiology and Behavior, 82*, 109–114.

Beck, J., & Shaw, W. A. (1961). The scaling of pitch. *American Journal of Psychology, 74*, 242–251.

Birnbaum, M. H. (1980). Comparison of two theories of "ratio" and "difference" judgments. *Journal of Experimental Psychology, 109*, 304–319.

Birnbaum, M. H. (1982). Problems with so-called "direct" scaling. In J. T. Kuznicki, R. A. Johnson, & A. F. Rutkiewic (Eds.), *Selected sensory methods: Problems and approaches to hedonics* (pp. 34–48). Philadelphia, PA: American Society for Testing and Materials.

Borg, G. (1961). Interindividual scaling and perception of muscular force. *Kungl Fysiografiska Sallskapets I Lund Forhandlingar, 31*, 117–125.

Borg, G. (1970). Relative response and stimulus scales. *Reports from the Institute of Applied Psychology* (Vol. 1, pp. 1–8). The University of Stockholm.

Borg, G. (1971). Psychological and physiological studies of physical work. In T. W. Singleton, J. G. Fox, & D. Whitfield (Eds.), *Measurement of man at work* (pp. 121–128). London: Taylor and Francis.

Borg, G. (1982). A category scale with ratio properties for intermodal and interindividual comparisons. In H. G. Geissler & P. Petzold (Eds.), *Psychophysical judgment and the process of perception* (pp. 25–34). New York, NY: North-Holland Publishing Company.

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

Borg, G. (1994). Psychophysical scaling: an overview. In J. Boivie, P. Hansson, & U. Lindblom (Eds.), *Touch, temperature, and pain in health and disease: Mechanisms and assessments*. Seattle, WA: IASP Press.

Borg, G., & Borg, P. (1987). On the relations between category scales and ratio scales and a method for scale transformation. In *Reports from the Institute of Applied Psychology* (pp. 1–14). University of Stockholm.

Borg, G., Diamant, H., Strom, L., & Zotterman, Y. (1967). The relation between neural and perceptual intensity: A comparative study on the neural and psychophysical response to task stimuli. *Journal of Physiology, 192*, 13–20.

Borg, G., & Hosman, J. (1970). The metric properties of adverbs. *Reports from the Institute of Applied Psychology* (Vol. 7, pp. 1–7). University of Stockholm.

Borg, G., & Lindblad, I. (1976). The determination of subjective intensities in verbal descriptions of symptoms. *Reports from the Institute of Applied Psychology* (Vol. 75, pp. 1–22). University of Stockholm.

Borg, G., Ljunggren, G., & Marks, L. E. (1985). General differential aspects of perceived exertion and loudness assessed by two new methods. *Reports from the Institute of Applied Psychology* (Vol. 636, pp. 1–13). University of Stockholm.

Boring, E. G. (1929). *Sensation and perception in the history of experimental psychology*. New York, NY: Appleton-Century-Crofts.

Cain, W. S. (1976). Olfaction and common chemical sense: Some psychophysical contrasts. *Sensory Processes, 1*, 57–67.

Cardello, A. V., Lawless, H. T., & Schutz, H. G. (2008). Effects of extreme anchors and interior label spacing on labeled affective magnitude scales. *Food Quality and Preference, 19*, 473–480.

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

Cordonnier, S. M., & 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, D. N., & Evans, G. (2008). Construction and validation of a psychometric scale to measure consumers' fears of novel food technologies: The food technology neophobia scale. *Food Quality and Preference, 19*, 704–710.

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

Daning, R. (1983). Intraindividual consistencies in cross-modal matching across several continua. *Perception and Psychophysics, 33*, 516–522.

Dawson, W. E. (1971). Magnitude estimation of apparent sums and differences. *Perception and Psychophysics, 9*, 368–374.

de Araujo, I. E., Rolls, E. T., Kringelbach, M. L., McGlone, F., & Phillips, N. (2003). Taste-olfactory convergence, and the representation of the pleasantness of flavour, in the human brain. *European Journal of Neuroscience, 18*(7), 2059–2068.

Diamond, J., & Lawless, H. T. (2001). Context effects and reference standards with magnitude estimation and the labeled magnitude scale. *Journal of Sensory Studies, 16*, 1–10.
Ekman, G. (1964). Is the power law a special case of Fechner's law? *Perceptual and Motor Skills, 19*, 730.
El Dine, A. N., & 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*(2), S97–S106.
Engen, T., & McBurney, D. H. (1964). Magnitude and category scales of the pleasantness of odors. *Journal of Experimental Psychlogy, 68*, 435–440.
Eriksen, C. W., & Hake, H. W. (1957). Anchor effects in absolute judgments. *Journal of Experimental Psychology, 53*(2), 132–138.
Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig, Germany: Breitkopf und Hartel.
Fechner, G. T. (1898). *Vorschule der aesthetik*. II. Leipzig: Breitkopf & Hartel.
Finn, A., & 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*, 12–15.
Flynn, T. N., Louviere, J. J., Peters, T. J., & Coast, J. (2007). Best-worst scaling: What it can do for health care research and hot to do it. *Journal of Health Economics, 26*, 171–189.
Freyd, M. (1923). The graphic rating scale. *Journal of Educational Psychology, 14*, 83–102.
Galanter, E. (1966). *Textbook of elementary psychology*. San Francisco: Holden-Day.
Gay, C., & Mead, R. (1992). A statistical appraisal of the problem of sensory measurement. *Journal of Sensory Studies, 7*, 205–228.
Gescheider, G. A. (1988). Psychophysical scaling. *Annual Review of Psychology, 39*, 169–200.
Gescheider, G. A., & Joelson, J. M. (1983). Vibrotactile temporal summation for threshold and suprathreshold level of stimulation. *Perception and Psychophysics, 33*, 156–162.
Giovanni, M. E., & Pangborn, R. M. (1983). Measurement of taste intensity and degree of liking of beverages by graphic scales and magnitude estimation. *Journal of Food Science, 48*, 1175–1182.
Green, B. G., Dalton, P., Cowart, B., Shaffer, G. S., Rankin, K., & Higgins, J. (1996). Evaluating the 'Labeled Magnitude Scale' for measuring sensations of taste and smell. *Chemical Senses, 21*, 323–334.
Green, B. G., & Lim, J. (2009). Evidence that repeated threshold testing can alter the perceived intensity of taste. *Chemical Senses, 34*, A120.
Green, B. G., Shaffer, G. S., & Gilmore, M. M. (1993). Derivation and evaluation of a semantic scale of oral sensation magnitude with apparent ratio properties. *Chemical Senses, 18*, 683–702.
Greene, J. L., Bratka, K. J., Drake, M. A., & 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., Prajapati, R., & McGlone, F. (2007). Labeled magnitude scales for oral sensations of sweetness, dryness, pleasantness and unpleasantness. *Food Quality and Preference, 18*, 342–352.
Guilford, J. P. (1936). *Psychometric methods*. New York: McGraw-Hill.
Hein, K. A., Jaeger, S. R., Carr, B. T., & Delahunty, C. M. (2008). Comparison of five common acceptance and preference methods. *Food Quality and Preference, 19*, 651–661.
Hellman, R. P. (1976). Growth of loudness at 1000 and 3000 Hz. *Journal of the Acoustical Society of America, 60*, 672–679.
Hellman, R. P., & Zwislocki, J. (1963). Monaural loudness function at 1000cps and interaural summation. *Journal of the Acoustical Society of America, 35*(6), 856–865.
Hellman, R. P., & Zwislocki, J. J. (1964). Loudness function of a 1000 cps tone in the presence of a masking noise. *Journal of the Acoustical Society of America, 36*, 1618–1627.
Helson, H. (1947). Adaptation-level as frame of reference for prediction of psychophysical data. *American Journal of Psychology, 1*, 29.
Helson, H. (1948). Adaptation-level as a basis for a quantitative theory of frames of reference. *Psychological Review, 55*(6), 297–313.
Henion, K. E. (1971). Odor pleasantness and intensity-single dimension. *Journal of Experimental Psychology, 90*(2), 275.
Hollingworth, H. L. (1910). The central tendency of judgment. *Journal of Philosophical and Psychological Science Methods, 7*, 461–469.
Hosman, J., & Borg, G. (1970). The mean and standard deviation of cross-modality matches: A study of individual scaling behavior. *Reports from the Institute of Applied Psychology* (Vol. 3). University of Stockholm.
Hulshoff Pol, H. E., Hijman, R., Baare, W. F., & van Ree, J. M. (1998). Effects of context on judgements of odor intensities in humans. *Chemical Senses, 23*(2), 131–135.
Jaeger, S. R., & 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.
Jaeger, S. R., Jorgensen, A. S., Aaslyng, M. D., & 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.
Jones, F. N. (1974). Overview of psychophysical scaling. In E. C. Carterette & M. P. Friedman (Eds.). *Handbook of perception: Psychophysical judgment and measurement* (Vol. II, pp. 343–360). New York, NY: Academic Press Inc..
Jones, L. V., Peryam, D. R., & Thurstone, L. L. (1955). Development of a scale for measuring soldiers' food preferences. *Food Research, 20*, 512–520.
Jones, L. V., & Thurstone, L. L. (1955). The psychophysics of semantics: An experimental investigation. *Journal of Applied Psychology, 39*(1), 31–36.
Kamenetzky, J. (1959). Contrast and convergence effects in ratings of foods. *Journal of Applied Physiology, 43*, 47–52.
Kim, K.-O., & O'Mahony, M. (1998). A new approach to category scales of intensity I: Traditional versus rank-rating. *Journal of Sensory Studies, 13*, 241–249.
Koo, T.-Y., Kim, K.-O., & 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.
Lawless, H. T. (1983). Contextual effects in category ratings. *Journal of Testing Evaluation, 11*, 346–349.
Lawless, H. T. (2010). Commentary on "Comparative performance of the nine-point hedonic, hybrid and self-adjusting scales in the generation of internal preference maps". *Food Quality and Preference, 21*, 165–166.
Lawless, H. T., & Heymann, H. (1998). *Sensory evaluation of food: principles and practices*. New York: Chapman & Hall.
Lawless, H. T., Horne, J., & Spiers, W. (2000). Contrast and range effects for category, magnitude and labeled magnitude scales in judgements of sweetness intensity. *Chemical Senses, 25*, 85–92.
Lawless, H. T., & Malone, G. J. (1986a). The discriminative efficiency of common scaling methods. *Journal of Sensory Studies, 1*(1), 85–98.
Lawless, H. T., & Malone, G. J. (1986b). A comparison of rating scales: sensitivity, replicates and relative measurement. *Journal of Sensory Studies, 1*(2), 155–174.
Lawless, H. T., Popper, R., & Kroll, B. J. (2010). A comparison of the labeled magnitude (LAM) scale, an 11-point category sale and the traditional nine-point hedonic scale. *Food Quality and Preference, 2*, 4–12.
Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychiatric, 22*, 1–55.
Lim, J., & Fujimaru, T. (2010). Evaluation of the labeled hedonic scale under different experimental conditions. *Food Quality and Preference, 21*, 521–530.
Lim, J., Wood, A., & Green, B. G. (2009). Derivation and evaluation of a labeled hedonic scale. *Chemical Senses, 34*, 739–751.
MacKay, D. M. (1963). Psychophysics of perceived intensity: A theoretical basis for Fechner's and Stevens' Laws. *Science, 139*, 1213–1216.
Marchisano, C., Lim, J., Cho, H. S., Suh, D. S., Jeon, S. Y., Kim, K. O., et al. (2003). Consumers report preference when they should not: A cross-cultural study. *Journal of Sensory Studies, 18*, 487–516.
Marks, L. E. (1966). Brightness as a function of retinal locus. *Perception and Psychophysics, 1*, 335–341.
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. (1979). Summation of vibrotactile intensity: An analog to auditory critical bands? *Sensory Processes, 3*, 188–203.
Marks, L. E. (1992). The contingency of perceptual processing: context modifies equal-loudness relations. *Psychological Science, 3*, 285–291.
Marks, L. E., & Bartoshuk, L. M. (1979). Ratio scaling of taste intensity by a matching procedure. *Perception and Psychophysics, 26*, 335–339.
Marks, L. E., Borg, G., & Ljunggren, G. (1983). Individual differences in perceived exertion assessed by two new methods. *Perception and Psychophysics, 34*, 280–288.
Marks, L. E., & Warner, E. (1991). Slippery context effect and critical bands. *Journal of Experimental Psychology: Human Perception and Performance, 17*(4), 986–996.
Marley, A. A. J., & Louviere, J. J. (2005). Some probabilistic models of best, worst, and best-worst choices. *Journal of Mathematical Psychology, 49*, 464–480.
Mattes, R. D., & Lawless, H. T. (1985). An adjustment error in optimization of taste intensity. *Appetite, 6*(2), 103–114.
McDaniel, M. R., & Sawyer, F. M. (1981). Preference testing of whiskey sour formulation: Magnitude estimation versus the 9-point hedonic scale. *Journal of Food Science, 46*, 182–185.
Mead, R., & Gay, C. (1995). Sequential design of sensory trials. *Food Quality and Preference, 6*, 271–280.
Mellers, B. A. (1983a). Evidence against "absolute" scaling. *Perception and Psychophysics, 33*, 523–526.
Mellers, B. A. (1983b). Reply to Zwislocki's views on "absolute" scaling. *Perception and Psychophysics, 34*, 405–408.
Mellers, B. A., & Birnbaum, M. H. (1982). Loci of contextual effects in judgment. *Journal of Experimental Psychology: Human Perception and Performance, 8*, 582–601.
Moskowitz, H. R. (1971). The sweetness and pleasantness of sugars. *American Journal of Psychology, 84*, 387–405.
Moskowitz, H. R. (1977). Magnitude estimation: Notes on what, how, when, and why to use it. *Journal of Food Quality, 3*, 195–227.
Moskowitz, H. R. (1980). Psychometric evaluation of food preferences. *Journal of Foodservice Systems, 1*, 149–167.
Moskowitz, H. R. (1982). Utilitarian benefits of magnitude estimation scaling for testing product acceptability. In J. T. kuznicki, R. A. Johnson, & A. F. Rutkiewic (Eds.), *Selected sensory methods: problems and approaches to measuring hedonics, ASTM STP 773*. Philadelphia, PA: American society for testing and materials.
Moskowitz, H. R. (2005). Thoughts on subjective measurement, sensory metrics and usefulness of outcomes. *Journal of Sensory Studies, 20*, 347–362.
Moskowitz, H. R., Dravnieks, A. L., & Klarman, L. (1976). Odor intensity and pleasantness for a diverse set of odorants. *Perception and Psychophysics, 9*, 122.
Moskowitz, H. R., Kluter, R. A., Westerling, J., & Jacobs, H. L. (1974). Sugar sweetness and pleasantness: evidence for different psychological laws. *Science, 184*(136), 583–585.
Moskowitz, H. R., & Sidel, J. L. (1971). Magnitude and hedonic scales of food acceptability. *Journal of Food Science, 36*, 677–680.

Moskowitz, H. R., Wolfe, K., & Beck, C. (1978). Sweetness and acceptance optimization in cola flavored beverages using combinations of artifical sweeteners – A psychophysical approach. *Journal of Food Quality, 2*, 17–26.
O'Mahony, M. (1982). Some assumptions and difficulties with common statistics for sensory analysis. *Food Technology, 36*(11), 75–82.
O'Mahony, M., & Heintz, C. (1981). Direct magnitude estimation of salt taste intensity with continuous correction for salivary adaptation. *Chemical Senses, 6*, 101–112.
O'Mahony, M., Masuoka, S., & Ishii, R. (1994). A theoretical note on difference tests: models, paradoxes and cognitive strategies. *Journal of Sensory Studies, 9*, 247–272.
O'Mahony, M., Park, H., Park, J. Y., & Kim, K. O. (2004). Comparison of the statistical analysis of hedonic data using analysis of variance and multiple comparisons versus an R-index analysis of ten ranked data. *Journal of Sensory Studies, 19*, 519–529.
Ohzawa, I., Sclar, G., & Freeman, R. D. (1982). Contrast gain control in the cat visual cortex. *Nature, 298*(5871), 266–268.
Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*(6), 407–418.
Parducci, A. (1974). Contextual effects: A range-frequency analysis. In E. C. Carterette & M. P. Friedman (Eds.). *Handbook of perception: Psychophysical judgment and measurement* (Vol. II, pp. 127–141). New York, NY: Academic Press Inc..
Parducci, A. (1982). Scale values and phenomenal experience: There is no psychophysical law! In H. G. Geissler & P. Petzold (Eds.), *Psychophysical judgment and the process of perception* (pp. 11–16). New York, NY: North-Holland Publishing Company.
Parducci, A., & Perrett, L. F. (1971). Category rating scales: effects of relative spacing and frequency of stimulus values. *Journal of Experimental Psychology Monography, 89*, 427–452.
Parducci, A., & Wedell, D. H. (1986). The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 496–516.
Parker, S., Murphy, D. R., & Schneider, B. A. (2002). Top-down gain control in the auditory system: Evidence from identification and discrimination experiments. *Perception and Psychophysics, 64*(4), 598–615.
Parker, S., & Schneider, B. (1994). The stimulus range effect: evidence for top-down control of sensory intensity in audition. *Perception and Psychophysics, 56*(1), 1–11.
Pearce, J. H., Korth, B., & Warren, C. B. (1986). Evaluation of three scaling methods for hedonics. *Journal of Sensory Studies, 1*, 27–46.
Peryam, D. R. (1950). *Problem of preference gets GM focus*. December: Food Industries.
Peryam, D. R., & Girardot, N. F. (1952). Advanced taste-test method. *Food Engineering, 24*, 58–61.
Peryam, D. R., & Pilgrim, F. J. (1957). Hedonic scale method of measuring food preference. *Food Technology, 11*, 9–14.
Poulton, E. C. (1979). Models for biases in judging sensory magnitude. *Psychological Bulletin, 86*, 777–803.
Prescott, J. (2009). Rating a new hedonic scale: a commentary on ''derivation and evaluation of a labeled hedonic scale'' by Lim, Wood and Green. *Chemical Senses, 34*(9), 735–737.
Rankin, K. M., & Marks, L. E. (1991). Differential context effects in taste perception. *Chemical Senses, 17*, 617–629.
Riskey, D. R. (1982). Effects of context and interstimulus procedures in judgments of saltiness and pleasantness. In J. T. Kuznicki, R. A. Johnson, & A. F. Rutkiewic (Eds.), *Selected sensory methods: Problems and approaches to measuring hedonics* (pp. 71–83). Am Soc Testing and Mat.
Riskey, D. R., Parducci, A., & Beauchamp, G. K. (1979). Effects of context in judgements of sweetness and pleasantness. *Perception and Psychophysics, 26*, 171–176.
Rolls, E. T., Kringelbach, M. L., & de Araujo, I. E. (2003). Different representations of pleasant and unpleasant odours in the human brain. *European Journal of Neuroscience, 18*(3), 695–703.
Savage, C. W. (1970). *The measurement of sensation*. Berkeley, CA: University of California Press.
Schifferstein, H. N. J. (1995). Contextual shifts in hedonic judgments. *Journal of Sensory Studies, 10*, 381–392.
Schifferstein, H. N. J., & Frijters, A. E. R. (1992). Contextual and sequential effects on judgments of sweetness intensity. *Perception and Psychophysics, 52*, 243–255.
Schneider, B., & Parker, S. (1990). Does stimulus context affect loudness or only loudness judgments? *Perception and Psychophysics, 48*(5), 409–418.
Schneider, B., Parker, S., & Moraglia, G. (1996). The effect of stimulus range on perceived contrast: evidence for contrast gain control. *Canadian Journal of Experimental Psychology, 50*(4), 347–355.
Schutz, H. G., & Cardello, A. V. (2001). A labeled affective magnitude (LAM) scale for assessing food liking/disliking. *Journal of Sensory Studies, 16*, 117–159.
Shand, P. J., Hawrysh, Z. J., Hardin, R. T., & Jeremiah, L. E. (1985). Descriptive sensory assessment of beef steaks by category scaling, line scaling and magnitude estimation. *Journal of Food Science, 50*, 495–500.
Small, D. M., Gregory, M. D., Mak, Y. E., Gitelman, D., Mesulam, M. M., & Parrish, T. (2003). Dissociation of neural representation of intensity and affective valuation in human gustation. *Neuron, 39*(4), 701–711.
Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: John Wiley and Sons.
Stevens, S. S. (1953). On the brightness of lights and loudness of sounds. *Science, 118*, 576.
Stevens, S. S. (1955). The measurement of loudness. *Journal of the Acoustical Society of America, 27*, 815–829.
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*(3), 153–181.
Stevens, S. S. (1971). Issues in psychophysical measurement. *Psychological Review, 78*, 426–450.
Stevens, S. S. (1974). Perceptual magnitude and its measurement. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception: Psychophysical judgment and measurement* (pp. 361–389). New York, NY: Academic Press, Inc.
Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York: Wiley.
Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology, 54*(6), 377–411.
Stone, H., Sidel, J., Oliver, S., Woolsey, A., & Singleton, R. C. (1974). Sensory evaluation by quantitative descriptive analysis. *Food Technology, 28*, 24–34.
Teghtsoonian, R. (1971). On the exponents in Stevens' law and the constant in Ekman's law. *Psychological Review, 78*, 71–80.
Teghtsoonian, R. (1973). Range effects in psychophysical scaling and a revision of Stevens' law. *American Journal of Psychology, 86*, 3–27.
Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review, 34*, 273–286.
Thurstone, L. L. (1927b). Psychophysical analysis. *American Journal of Psychology, 38*, 368–389.
Torgerson, W. S. (1961). Distances and ratios in psychophysical scaling. *Acta Psychologica, 19*, 201–205.
Treisman, M. (1964). Sensory scaling and the psychophysical law. *Quarterly Journal of Experimental Psychology, 16*, 11–12.
Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91*, 68–111.
Verrillo, R. T., Fraioli, A. J., & Smith, R. L. (1969). Sensory magnitude of vibrotactile stimuli. *Perception and Psychophysics, 6*, 366–372.
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., & Da Silva, M. A. A. P. (2009). Comparative 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., & Da Silva, M. A. A. P. (2000). Performance of three affective methods and diagnosis of the ANOVa model. *Food Quality and Preference, 11*, 363–370.
Villanueva, N. D. M., Petenate, A. J., & Da Silva, M. A. A. P. (2005). Performance of the hybrid hedonic scale as compared to the traditional hedonic, self-adjusting and ranking scales. *Food Quality and Preference, 16*, 691–703.
Villegas-Ruiz, X., Angulo, O., & O'Mahony, M. (2008). Hidden and false ''preferences'' on the structured 9-point hedonic scale. *Journal of Sensory Studies, 23*, 780–790.
Ward, L. M. (1991). Associative measurement of psychological magnitude. In S. J. Bolanowski & G. A. Gescheider (Eds.), *Ratio scaling of psychophysical magnitude: In honor of the memory of S.S. Stevens*. Hillsdale, NJ.: Lawrence Erlbaum.
Warren, C. B. (1981). Development of fragrances with functional properties by quantitative measurement of sensory and physical parameters, ''Odor Quality and Chemical Structure''. In *ACS Symposium Series 148* (pp. 57–77). Washington, DC: ACS.
Winston, J. S., Gottfried, J. A., Kilner, J. M., & Dolan, R. J. (2005). Integrated neural representations of odor intensity and affective valence in human amygdala. *Journal of Neuroscience, 25*(39), 8903–8907.
Yao, E., Lim, J., Tamaki, K., Ishii, R., Kim, K.-O., & O'Mahony, M. (2003). Structured and unstructured 9-point hedonic scales: A cross-cultural study with American, Japanese and Korean consumers. *Journal of Sensory Studies, 18*, 115–139.
Yeh, L. L., Kim, K.-O., Chompreeda, P., Rimkeeree, H., Yau, N. J. N., & Lundahl, D. S. (1998). Comparison in use of the 9-poin hedonic scale between Americans, Chinese, Koreans, and Thai. *Food Quality and Preference, 9*(6), 413–419.
Zealley, A. K., & Aitken, R. C. (1969). Measurement of mood. *Proceedings of the Royal Society for Medicine, 62*(10), 993–996.
Zellner, D. A., Allen, D., Henley, M., & Parker, S. (2006). Hedonic contrast and condensation: good stimuli make mediocre stimuli less good and less different. *Psychonomic Bulletin and Review, 13*(2), 235–239.
Zellner, D. A., Kern, B. B., & Parker, S. (2002). Protection for the good: subcategorization reduces hedonic contrast. *Appetite, 38*, 175–180.
Zellner, D. A., Rohm, E. A., Bassetti, T. L., & Parker, S. (2003). Compared to what? Effects of categorization on hedonic contrast. *Psychonomic Bulletin and Review, 10*(2), 468–473.
Zwislocki, J. J. (1983a). Absolute and other scales: Question and validity. *Perception and Psychophysics, 33*, 593–594.
Zwislocki, J. J. (1983b). Group and individual relations between sensation magnitude and their numerical estimates. *Perception and Psychophysics, 33*, 460–468.
Zwislocki, J. J., & Goodman, D. A. (1980). Absolute scaling of sensory magnitudes: A validation. *Perception and Psychophysics, 28*, 28–38.