Chapter 14

Acceptance Testing

Abstract An alternative to choice procedures for assessing the consumer appeal of foods is to use a rating scale for the degree of liking or disliking, otherwise known as acceptability scaling or acceptance testing. This chapter illustrates procedures for acceptability scaling, starting with the traditional 9-point hedonic scale in widespread use. Alternative types of acceptance scales are shown. The just-about-right (JAR) scale is illustrated and its statistical analyses are discussed.

About 1930, Dr. Beebe-Center, psychologist at Harvard, wrote a book in which he reported the results of investigations of the pleasantness/unpleasantness of dilute solutions of sucrose and sodium chloride. He called his measurements hedonics. I liked the word, which is both historically accurate and now well installed, and used it in the first official report on the new scale.

—David Peryam, "Reflections" (1989)

Contents

14.1 Introduction: Scaled Liking Versus Choice

The previous chapter dealt with consumer tests involving choices among alternatives and ranking of alternative products. In this chapter we will look at methods for scaling the degree of acceptability of foods. Note that these methods do not require a choice between alternatives. In theory, acceptance scaling can be done on a single product, although a one-product test is usually not very informative and lacks any baseline for making comparisons. A scale that measures the sensory appeal of a product has distinct advantages over a simple choice task. Most importantly, it provides some information on whether the product is liked or disliked in some absolute sense. In a preference test, I might dislike both products and choose the least offensive. In such a case it would obviously not be a good idea to produce or try to sell either version of the product, but the preference test does not tell you this fact. In

H.T. Lawless, H. Heymann, *Sensory Evaluation of Food*, Food Science Text Series,
DOI 10.1007/978-1-4419-6488-5_14,

addition to the liking or disliking information, preference can be inferred from a superior acceptance score of one product over another. For these reasons, many sensory professionals consider acceptance tests to be a better choice than a preference test. Of course, there is no rule against asking both kinds of questions in a test with multiple products and this is often done in consumer field tests as discussed in Chapter 15.

Acceptability data from scales are useful for a number of additional purposes. It is also possible to convert the hedonic scale results to paired preference or rank data (Rohm and Raaber, 1991). Since the scaled acceptance data are "richer" in information, it is possible to derive these other simpler measures from them. Hedonic data can be used in preference mapping techniques (for examples, see Greenhoff and MacFie, 1994; Helgensen et al., 1997; McEwan, 1996). This is a useful technique that allows visualization of the directions for product preferences in spatial models of a product set (see Chapter 19). In spatial models from multivariate analyses, products are represented by points in the space and products that are similar are positioned close together. Dimensions or attributes that differentiate the products can be inferred from product positions, from opposites positioned at different sides, and from interpretation of the axes of the space. Preferences of individual consumers can be projected as vectors through the space to show directions of increased liking. These vectors can then suggest directions for product optimization. Also, differences in the preferred directions for different consumers can help discover market segments or groups with different likes and dislikes.

14.2 Hedonic Scaling: Quantification of Acceptability

The most common hedonic scale is the 9-point hedonic scale shown in Fig. 14.1. This is also known as a degree of liking scale. This scale has achieved wide popularity since it was first invented in the 1940s at the Food Research Division of the Quartermaster Food and Container Institute in Chicago, Illinois (Peryam and Girardot, 1952). David Peryam coined the name hedonic scale for the 9-point scale used to determine degree of liking for food products (Peryam and Girardot, 1952). The hedonic scale assumes consumer preferences exist on a continuum and that preference can be categorized by responses based on like and dislike. The scientists at the Quartermaster Institute evaluated the scale using soldiers in the field, in the laboratory, and in attitude surveys (Peryam and Pilgrim, 1957). Samples were served to panelists monadically (one at a time) and the panelists were asked to indicate their hedonic response to the sample on the scale. Research at the Quartermaster Institute had indicated that the specific way the scale appeared on the score sheet, whether the scale was printed horizontally or vertically, or whether the like or dislike side was encountered first, did not affect results. Jones et al. (1955) found that the ideal number of categories was 9 or 11 and the researchers at the University of Chicago and the Quartermaster Institute decided to use the 9-point version, because it fits better on the typing paper of that era (Peryam, 1989).

LIKE EXTREMELY

LIKE VERY MUCH

LIKE MODERATELY

LIKE SLIGHTLY

NEITHER LIKE NOR DISLIKE

DISLIKE SLIGHTLY

DISLIKE MODERATELY

DISLIKE VERY MUCH

DISLIKE EXTREMELY

Fig. 14.1 The phrases for the 9-point hedonic scale for food acceptance testing. Responses on this scale are usually assigned values from 1 to 9, 1 for dislike extremely and 9 for like extremely.

> Why does the hedonic scale have nine categories, rather than more or less? Economy perhaps? Preliminary investigation had shown that discrimination between foods and reliability tended to increase up to eleven categories, but we encountered, in addition to the dearth of appropriate adverbs, a mechanical problem due to equipment limitations. Official government paper was only 8″ wide and we found that typing eleven categories horizontally was not possible. So we sacrificed a theoretical modicum of precision for a real improvement in efficiency at the moment. p. 23

The words chosen for each scale option were based on approximately equal differences as determined by Thurstonian methods (see Chapter 7). Thus the scale,

psychologically, has ruler-like properties that are not necessarily present in other less carefully constructed liking scales. This equal-interval property is important in the assignment of numerical values to the response choices and to the use of parametric statistics in analysis of the data. Thus the sensory scientist should be cautious and avoid "tinkering" with the scale alternatives. It is important to resist pressure from non-sensory specialists or managers familiar with other scales to modify the scale or adopt alternatives.

The 9-point scale is very simple to use and is easy to implement. It has been widely studied and has been shown to be useful in the hedonic assessment of foods, beverages, and non-food products for decades. The US military has studied its applicability, validity, and reliability and the positive aspects of this scale have been widely accepted. Peryam and Pilgrim (1957) note that the hedonic rating can be affected by changes in environmental conditions (for instance, under field conditions versus cafeteria conditions) but the relative order of sample preference was usually not affected. In other words, the absolute magnitude of the hedonic score may increase or decrease but all samples had similar relative changes. Tepper et al. (1994) showed that consumers rank ordered and hedonically scored products similarly. It has been reported that the scale is reliable and has a high stability of response that is independent of region and to some extent of panel size. However, the applicability of the scale in other languages and cultures have not been as widely studied and it should be used cautiously in these situations.

The 9-point scale has been criticized on several grounds. Moskowitz (1980) suggested that the 9-point hedonic scale has potential problems associated with category scales such as the categories are not quite equally spaced, the neutral ("neither like nor dislike") category makes the scale less efficient and consumers tend to avoid the extreme categories. However, the initial calibration work indicated that this particular category scale has nearly equal-interval spacing although direct scaling methods seem to indicate that the distance from neutral to the like/dislike slightly categories is smaller than the other intervals (Schutz and Cardello, 2001). The neutral response category is important as it is a valid reaction to the product for some participants. Although many scales show "end use avoidance," this serves as a warning to those who are tempted to truncate the scale to fewer than nine points. Truncating a scale to seven or five points may effectively reduce it to five or three useful categories since end-category avoidance may still come into play. This is one of the forms of "tinkering" to be avoided. The other temptation is to reduce the number of negative response options, often under the misplaced philosophy that the company does not make or test any really bad products. Due to some of these concerns, sensory researchers have used other scales for assessing liking, including various line scales and magnitude estimation, discussed further below.

A recent approach that is growing in popularity is a modification of the hedonic scale based on re-scaling of the word phrases using magnitude estimation and placing them on a line scale with the added end anchors, "greatest imaginable like" and "greatest imaginable dislike." This is the labeled affective magnitude scale or LAM scale discussed in detail later in this chapter (Schutz and Cardello, 2001). The scale development is based on the procedures for the labeled magnitude scale of Green and colleagues (see Chapter 7). Since the development of this scale, others have been developed using similar techniques, notably for oral pleasantness and wetness/dryness (Guest et al., 2007), clothing comfort (Cardello et al., 2003), taste pleasantness (Keskitalo et al., 2007), hedonics in general (Bartoshuk et al., 2006), and perceived satiety (Cardello et al., 2005).

14.3 Recommended Procedure

14.3.1 Steps

The procedure for conducting a scaled acceptance test is very similar to that for the simple paired preference test, except of course that the responses are required after each individual product, rather than each pair. The steps in conducting an acceptance test are shown in Table 14.1. Samples may be served one at a time, a response required after each sample and then the sample returned to the kitchen or prep area. Alternatively, the samples can be placed all on one tray, but this requires the panelist to match the correct test sample to the correct three-digit code on the questionnaire. This is usually done correctly but there are no guarantees. Therefore it is safest with truly naïve consumers to

Table 14.1 Steps in conducting an acceptance test

1. Obtain samples and confirm test purpose, details, timetable, and consumer qualifications (e.g., frequency of product usage) with client
2. Decide testing conditions (sample size, volume, temperature, etc.)
3. Write instructions to the panelists and construct ballot
4. Recruit potential consumers
5. Screen for product usage to qualify
6. Set up counterbalanced orders
7. Assign random three-digit codes and label sample cups/plates
8. Conduct test
9. Analyze results
10. Communicate results to client or end user

serve products one at a time and retrieve them after each response. A sample ballot for acceptance scaling is shown in Fig. 14.2.

14.3.2 Analysis

The data from the 9-point scales are assigned the values one through nine, nine usually being the "like extremely" category. They are then analyzed using parametric statistics, *t*-tests on means for two products, or analysis of variance followed by comparisons of means for more than two products. Even though the scale may not achieve a true interval level of measurement, the parametric approach is usually justified based on the larger sample size in a consumer test.

14.3.3 Replication

Acceptance tests do not commonly involve replicated tastings on the same products by the same consumers. However, there are several reasons to consider replication. The first is that it may provide some additional information. Byer and Saletan (1961) used repeated tests on beers (judges were blind to the replication) over several days to see if there were systematic increasing or decreasing liking for some beers as opposed to others. Second, a replication may greatly increase the discrimination of products once consumers have a better idea of the range of products to be evaluated. In a study by Hein et al. (2008), the increase in product discrimination was especially pronounced for the 9-point hedonic scale. Third, the first judgment by a consumer may not be well predictive of later behaviors (Koster et al., 2003). Finally, replication will allow one to reduce the effects of serving order, especially any advantage that might occur for the item in the first position (Hottenstein et al., 2008; Wakeling and MacFie, 1995).

14.4 Other Acceptance Scales

A number of other methods have been used to quantify consumer acceptance and this should not perhaps be surprising given the amount of consumer testing done by sensory evaluation personnel as well as marketing researchers. The 9-point scale itself has been modified in various ways in attempts to improve product discrimination. For example, Yao et al. (2003) found that an unstructured version of the 9-point scale (lacking the verbal labels) produced a somewhat wider range of scale usage among American and Japanese (but not Korean) consumers. In the early development and testing of various hedonic scales, Peryam (1989) noted that there could be room for expansion, especially at the positive end of the scale, stating, "An 8-point unbalanced scale with more 'like' than 'dislike' categories was shown to be somewhat better than the standard 9-point one, but only when dealing with relatively well-liked foods" (1989, p. 24). In the section below, we will briefly consider line scales, magnitude estimation, labeled magnitude scales, which are a combination of line marking and ratio-type scaling, and some relative scales which allow adjustment of previous ratings.

14.4.1 Line Scales

There are a number of studies where the panelists were asked to indicate their hedonic responses on

Acceptability Test
Braised Trake

Name_____ Date_____

Tester Number____ ____ ____ ____ Session Code ___ ___ ___

In a previous survey you indicated you are a consumer of BRAISED TRAKE. Please check an answer below that describes your recent consumption of this product.

In the last __3__ months, about how often have you eaten BRAISED TRAKE? (check one)
____ less than once a month
____ more than once a month but less than every week
____ once a week or more

Please rinse your mouth with water before starting.
Your can rinse at any time during the test if you need to.

Please taste the samples according to the number on each page.
Do NOT go back and re-taste the samples once you have turned the page.
If you have any questions, please ask the server now.

Check one phrase to indicate your overall opinion of the product.

Sample #___387___

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

PLEASE GO TO THE NEXT PAGE.

Fig. 14.2 A sample ballot for an acceptability test. The samples are evaluated on the 9-point balanced hedonic scale. Each subsequent page will have a new scale for a product with a new three-digit code. The order of evaluation is thus controlled by what is printed on each page, with randomization, rotation, or counterbalancing of orders. There is also a check on product usage frequency at the top of the page. This could be used to confirm that the panelists selected are still users of the product and are thus qualified to be in the test. Such confirmation is recommended for standing panels, such as employees or consumers chosen from a data bank, where the product-usage questionnaire might have been filled out some time in the past. In a consumer field test, the usage frequency would confirm what had been determined in a telephone screening or other recruiting interview (e.g., mall intercept).

unstructured line scales, sometimes anchored by like and dislike on each end (Hough, et al., 1992; Lawless, 1977; Rohm and Raaber, 1991). Line scales are sometimes referred to as visual analog scales (or VAS). That line scales would find some application in hedonics is not surprising, as they became the standard

scaling method for descriptive analysis in the 1970s, and their extension into acceptance scaling would seem logical. Recently, a version of the line scale with pips or markers equally spaced along the line has been studied by Villanueva and colleagues and found to compare favorably against the 9-point scale in terms of product differentiation and identification of consumer segments (Villanueva and Da Silva, 2009; Villanueva et al., 2005). However, on a statistical basis any advantages were slight (Lawless, 2009). A simplified version of the labeled affective magnitude scale (LAM scale) was used by Wright (2007) in which the end anchors "greatest imaginable like(/dislike)" were used instead of the usual "like (dislike) extremely." Some line scales are shown in Fig. 14.3.

14.4.2 Magnitude Estimation

As discussed in Chapter 7, magnitude estimation is a scaling procedure in which people can use any numbers they wish and are asked to consider the ratios or proportions between products. In the case of acceptance, they would be told to make a mark twice as far from the origin if the product is liked twice as much. In bipolar magnitude estimation, there are positive and negative numbers used to indicate likes and dislikes. This is not done in unipolar magnitude estimation in which products are only scaled as a distance from the bottom, which presumably represents a product which would not be liked at all. Given that people have likes and dislikes, a bipolar scale makes much more sense. In a study comparing the results of the 9-point hedonic scale to those obtained from a unipolar magnitude estimation scale and a bipolar magnitude estimation scale, Pearce et al. (1986) found that the three scales gave data that were very similar in terms of reliability, precision, and discrimination. However, the product category was fabric and the fabrics were evaluated by touching. It is possible that these results could have been different if a more fatiguing product category such as a tasted food was evaluated.

Magnitude estimation went through a period of some interest and has been used for evaluation of a number of food products (Lavanaka and Kamen, 1994; McDaniel and Sawyer, 1981). In an unusual combination of line scaling with ratio instructions, Lawless (1977) used a bipolar line scale with a zero or neutral point in the middle. Participants were instructed to consider ratios, for example, "if the next sample is liked twice as much, make a mark on the line twice as far from zero." The problem of having a bounded scale was circumvented by telling the subjects they could tape additional scales to the end of the strip of paper to extend the scale beyond the strip if needed.

These scales have not found much favor in industrial practice, in part because of the popularity of the 9-point scale and in part because of the complicated task of having consumers consider ratios of liking/disliking.

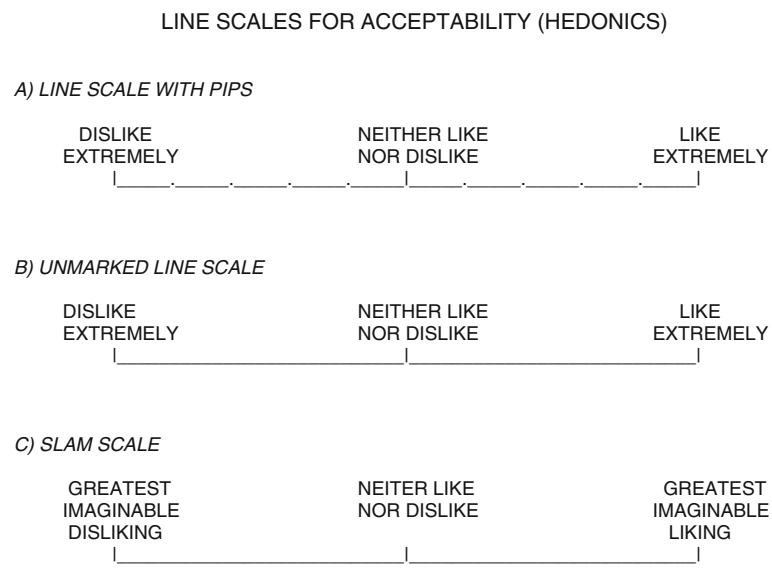


Fig. 14.3 Line scales for acceptability testing. (**a**) Line scale with small position markers, after Villanueva et al. (2005). (**b**) Unstructured line scale. (**c**) Simplified labeled affective magnitude scale (SLAM scale) after Wright (2007).

To make the process a little more user-friendly, other approaches have tried to simplify the task. In exploring the spacing of anchor terms for the LAM scale, a two-step process has been used, in which the magnitude of the hedonic reaction was considered, and then the sign, positive or negative, for the feeling evoked by the word (Cardello et al., 2008; Schutz and Cardello, 2001).

14.4.3 Labeled Magnitude Scales

The labeled affective magnitude scale (LAM, Fig. 14.4) was developed by Schutz and Cardello (2001) as an alternative to the commonly used 9-point category scale for measuring food acceptability (see also Cardello and Schutz, 2004). This scale was an extension of the procedure called the Labeled Magnitude Scale (LMS) that had been used for psychophysical intensity scaling. The LMS was developed by Green and colleagues and was based on earlier work by Borg for a hybrid "category–ratio scale" (Borg, 1982; Green et al., 1993). The LAM scale has been used for evaluation of consumer liking for teas (Chung and Vickers, 2007a, b) and in a comparison of young and older person's liking for different orange juices (Forde and Delahunty, 2004). The theoretical advantages to the LAM scale were proposed to be the following: First, because the word spacings were determined by magnitude estimation (ratio scaling instructions) one might presume that the data allow ratio-type conclusions ("Product A was liked twice as much as B."). In the published literature there are no examples in which this kind of conclusion has been drawn. Second, due to the high end anchors (greatest imaginable liking) people might have a similar idea of the intensity of reaction to this anchor (as proposed by Borg for his original intensity scale) and thus they would be working on the same or a similar psychological continuum.

Does the LAM scale provides any practical advantage over the traditional 9-point hedonic scale? The most important criterion for an advantage is whether one scale is better at finding differences among products (Lawless and Malone, 1986). In the original set of studies, performances of the LAM scale and the 9-point scale were similar (Schutz and Cardello, 2001). Two direct comparisons were conducted, one involving 51 food names and one involving 5 foods that were

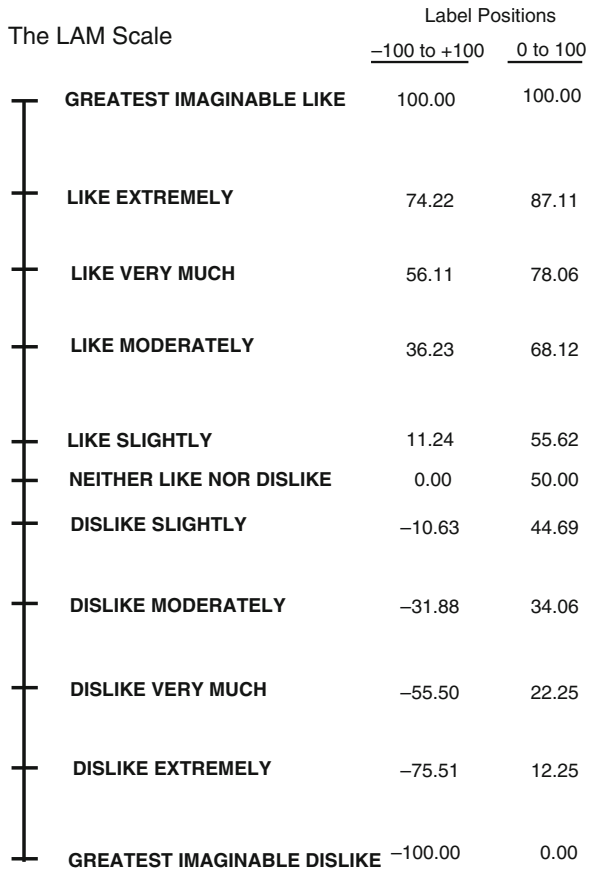


Fig. 14.4 The labeled affective magnitude (LAM) scale, from Schutz and Cardello (2001). For those who wish to construct the scale, the label positions are given on a 100-point and 200-point basis (–100 to +100), from Cardello and Schutz (2004).

actually tasted. Correlations between the mean values obtained on the two scales were +0.99 for the 51 food names and +0.98 for the tasted foods. Statistical differentiation was about the same in both cases. For the food names, there were about the same number of statistically different pairs of means (467 (LAM) versus 459 (9-point) out of 1,275 possible comparisons). A small advantage for the LAM scale was observed in comparing well-liked foods, i.e., those above the grand mean. The higher ends of the scale range were used more frequently with the LAM scale, consistent with the idea that it might be valuable for differentiating well-liked foods.

Several other studies have examined the performance of the two scales in direct comparisons. Greene et al. (2006) examined consumers' reactions to peanuts with and without fruity-fermented flavor defects. The

9-point hedonic scale only uncovered one significant pair of differences on one of the four scales, whereas the LAM scale showed four pairs of significant differences (out of 12 possible) and on three of the four scales. Rather than well-liked foods, these peanut samples scored very near the neutral point on both scales. El Dine and Olabi (2009) found that the LAM scale was as good and sometimes better than the 9-point scale in differentiating well-liked foods. However, in an extensive consumer study with several product categories, Lawless et al. (2009) found that in some cases the LAM was superior to the 9-point and in others the 9-point scale fared better. This was true for both product differentiation and correlation between the product that was best liked and the type of product the consumers said they most often purchased (a kind of validity check).

At this point it appears that the scales perform similarly, with a slight advantage to the LAM scale, which could be considered a viable alternative to the traditional 9-point scale, especially if well-liked foods are to be compared. There has been some discussion of whether the high end anchor should refer to the greatest imaginable like (dislike) for any kind of sensory experience (food or non-food) or whether the anchor should be more general or refer specifically to "foods like this." Using a more extreme end anchor (any imaginable sensory experience of any kind) will result in compression of the ratings toward the center of the scale, a context effect (Cardello et al., 2008). Because response compression is generally undesirable and it would be better to encourage fuller use of the scale, the choice of an extreme high end anchor (such as "greatest imaginable liking for any experience") is best avoided.

14.4.4 Pictorial Scales and Testing with Children

Hedonic scaling can also be achieved using face scales, frequently these are simple "smiley" faces (see Chapter 7) but they may also be more representational, involving animal cartoons (Moskowitz, 1986) or more realistic pictures of adults (Meilgaard et al., 1991). A variety of these pictorial scales can be found in Resurreccion (1998). These scales were invented for use by children or illiterate persons (Coetzee, 1996). However, young children may not have the cognitive skills to infer that the picture is supposed to indicate their internal responses to the product. Additionally, they may be distracted by the pictures. Kroll (1990) showed that verbal descriptors, the so-called P&K scale worked better with children than either the 9-point hedonic scale or facial scales. The terms in this scale are shown in Fig. 14.5. Kroll urged further exploration of the

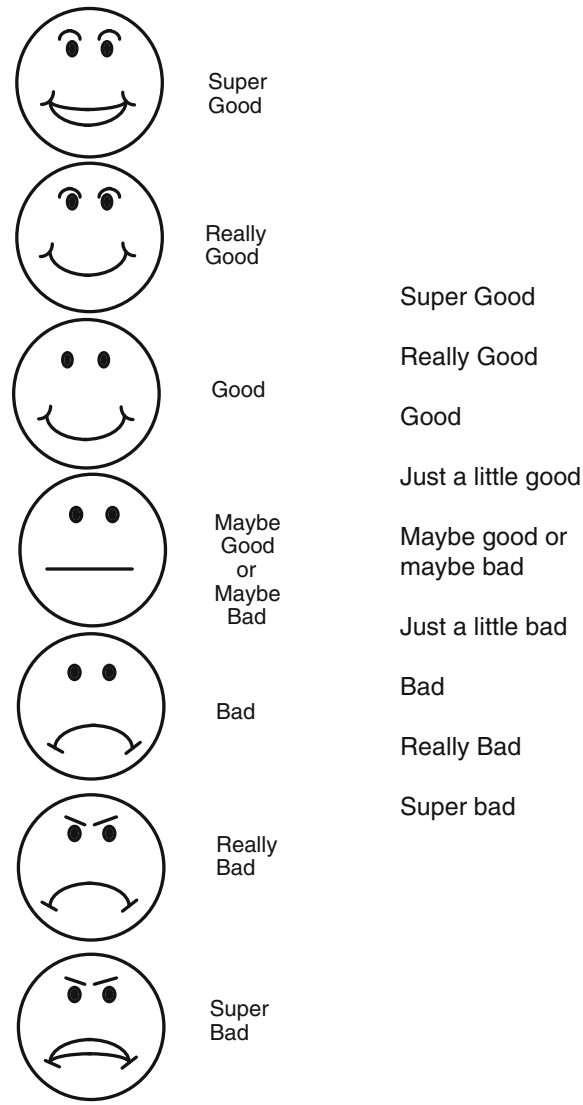


Fig. 14.5 Scales used with children. The *left side* shows an example of a facial scale constructed by Chen et al. (1996). The *right side* shows the super good–super bad verbal scale of Kroll (1990).

P&K scale with low-preference samples and with children under 5 years old (Schraidt, 1991). Chen et al. (1996) showed that 3-point facial hedonic scales with P&K verbal descriptors could be used with 36–47-month-old children, that a 5-point facial version could be used with 47–59-month-old children and that a 7-point version could be used with children 5 years and older. The facial scale used by Chen et al. is shown in Fig. 14.5. The facial scales have a long history of use in the study of food preferences and food habits among children (Birch 1979, Birch et al., 1980, 1982). Pagliarini et al. (2003) and Pagliarini et al. (2005) used an Italian version of the smile scale with verbal labels to study the acceptability of school lunch items and meal item combinations among Italian schoolchildren. Head et al. (1977) found that a 5-point scale (great, good, OK, bad, and terrible) was used reliably by elementary school children in grades 4–6 and secondary school children in grades 10–12.

An alternative to facial scales when testing children is to resort to simple paired preference. Kimmel et al. (1994) concluded that children as young as 2 years old could reliably perform a paired preference test if the appropriate environment and a one-on-one verbal test protocol were used. These authors also found that a 7-point facial hedonic scale anchored with words ranging from "super good" to "super bad" could be used consistently by children as young as 4 years old. Schmidt and Beauchamp (1988) also observed that 3-year-old children could reliably indicate their preference for odors using a paired test involving puppets. Bahn (1989) analyzed preference judgments for cereals made by 4- and 5-year-old children and by 8- and 9-year-old children using multidimensional scaling. Brand names had little effect on children's preferences and most preferences were based on sensory-affective responses to the cereals. Perceptual maps from the younger and older children were similar.

Preference or acceptance testing with children can be done with a few modifications from the adult methods. These often include the following: (1) one-on-one testing in most cases, to insure compliance, understanding, and to minimize social influences, (2) children can respond to either verbal scales or pictorial scales, (3) scales may need to be truncated for use with younger children, (4) paired preference testing is suitable for very young children in the ranges about 4–5 years. Below that age, likes and dislikes must be inferred from behaviors, such as counting oral contacts in an ad lib situation (Engen, 1974, 1978; Lawless et al., 1982–1983) or from ingestion or sucking (Engen et al., 1974).

14.4.5 Adjustable Scales

Two kinds of adjustable scales have appeared in the literature although to our knowledge they have not found wide acceptance in industrial practice. Gay and Mead proposed a method of scaling in which consumers would look at all the products to be scaled, and place the highest at the top of the scale, the lowest at the bottom, and then partition all of the others at appropriate intermediate marks on the scale (Gay and Mead, 1992; Mead and Gay, 1995), much like a ranking. The advantage of this method is that it eliminates differences among respondents in their choice of what scale range to use, as everyone uses the whole range of the scale. The disadvantage is that the scale is truly relative, i.e., no absolute information about degree of liking is obtained, only the relative positions of products. Although this can be applied to evaluating the perceived intensity of an attribute (like sweetness), and perhaps is most sensible for that purpose, it can be used for hedonics as well (Villanueva and Da Silva, 2009).

Another adjustable scaling method is the "rank rating" method (Kim and O'Mahony, 1998). In this method, the category scale is represented pictorially on the table in front of the consumer. Each product is tasted and the cup or sample is placed on the table in its appropriate category. As the participant proceeds through the test, they are allowed to change the position of products already rated. There are thus two important aspects of this procedure that could potentially enhance product discrimination. First, the consumer can see where the previous products were rated and second, they can change their minds. If the first product was placed too high or too low on the scale, relative to the position of the second product, the situation can be remedied. Whether this procedure is advantageous remains to be seen as it does not have an extensive record at this point. The initial experiments using intensity ratings of salt solutions showed fewer reversals, defined as a high concentration of NaCl being rated lower on the scale (Kim and O'Mahony, 1998). Comparisons to the 9-point hedonic scale have shown only small advantages (if any) to

the rank-rating procedure (Cordonnier and Delwiche, 2008; O'Mahony et al., 2004). Like the adjustable scale of Gay and Mead, rank rating may add some degree of relativity to the ratings (as opposed to product having absolute meanings regarding degree of liking). Both of these methods could be more susceptible to context effects and sequential dependencies (see Chapter 9).

14.5 Just-About-Right Scales

14.5.1 Description

A popular scale that combines intensity and hedonic judgments is the just-right or just-about-right (JAR) scales (Rothman and Parker, 2009). These scales, shown in Fig. 14.6 are bipolar, having opposite end-anchors and a center point. The end anchors are "Too little" and "Too much" of a specific attribute or a phrase such as "Too sweet" and "Not sweet enough." The center point can be labeled "just-right" but because it is felt that the choice of "just-right" may entail too strong a commitment on the part of the respondent, the center choice is usually rephrased as "just-about-right." The just-right scale is designed to measure the consumer's reaction to a specific attribute. For example, a just-right scale anchored with "Not salty enough" on the left, "Just-right" in the center and "Much too salty" on the right was used by Shepherd et al. (1989) to evaluate soups.

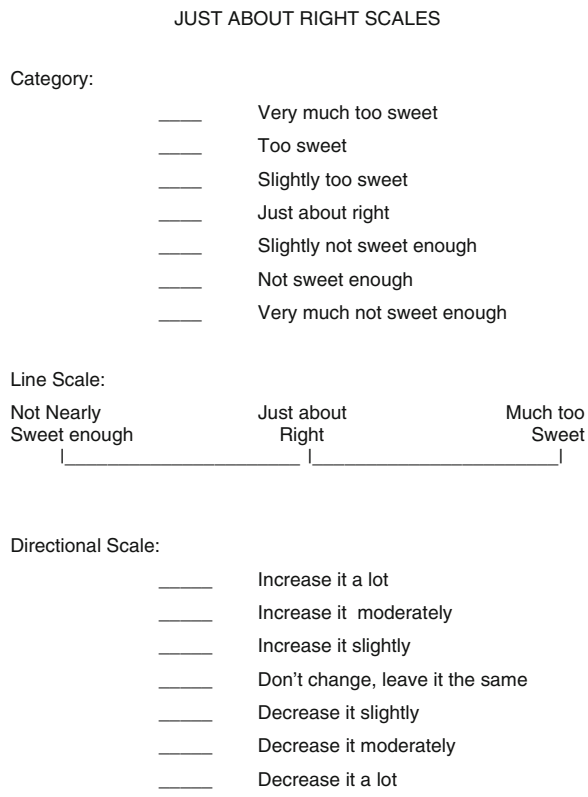


Fig. 14.6 Just-about-right scales. *Top*: a simple category scale. *Center*: an unstructured line scale. *Bottom*: a direction-of-change scale. Further examples can be found in Rothman and Parker (2009).

The just-right scales are popular for the direct information that they give on specific attributes to be optimized. Product developers like this information and so do managers. Furthermore, the concept of deviation from ideal taps into a basic decision when we react to products. For example, people commonly say that the coffee is too strong or too weak or that the wine is too sweet or too tannic. Whether we are aware of it or not, our opinions can be affected by what we expect and what we would like to obtain in terms of the sensory stimulation from a product. Many sensory continua have an optimum or "bliss point" (McBride, 1990). Booth has formulated a quantitative theory of food quality based upon the deviation of sensory attributes from their ideal levels (Booth, 1994, 1995). Using regular hedonic scaling, the "bliss point" appears as a peak in a non-monotonic function. The just-right data, however, "unfold" this function as shown in Fig. 14.7. Sometimes the unfolded function is linear or at least monotonic, leading to simpler modeling or curve fitting.

The most obvious application of this information is in optimization of a product's key attributes. Intensity and hedonic judgments are combined to provide directional information for product re-formulation. In a consumer test, this can be part of the final field test to insure that no gross errors have been made in the product formulation. The JAR scale gives information that can be diagnostic or explanatory if the overall product appeal is lacking. Earlier in the product development process, the scales can be useful in comparing different versions of the product. Another useful piece of information can be the identification of different segments of consumers who prefer different levels of a sensory attribute. When combined

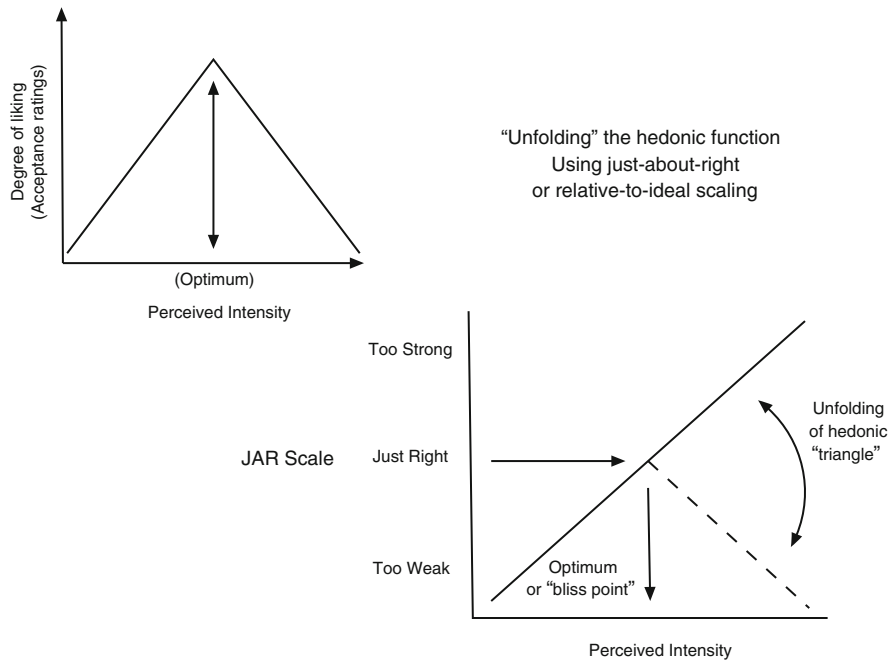


Fig. 14.7 "Unfolding" of the peaked hedonic function by the just-right scale. This can produce a linear relationship of just-right scores against sensory intensity or against ingredient concentrations (usually on a log scale). The slope of this line, relative to error, is indicative of the tolerance of the consumer for deviations from the ideal level.

with hedonic judgments, the potential impact of being off from the just-right point can be estimated using "penalty analysis" (discussed below). Another advantage is that these scales give directional information for change and can do so by testing only a single product (no complicated designed experiments are required). Some direct comparisons have been made concerning the performance of JAR scales versus other more traditional methods and no consistent advantage is apparent at this time (Bower and Boyd, 2002; Epler et al., 1997; Popper and Kroll, 2005). A good review of the issues involved can be found in Van Trijp et al. (2007).

14.5.2 Limitations

There are several concerns and pitfalls to be aware of when using JAR scales. The use of JAR scales assumes that all the consumers understand what the attribute listed on the score sheet is referring to. In other words, the consumers must have a common idea or consensus understanding of the attribute in question. This limits the just-right scale to a few simple attributes that are widely understood such as sweetness and saltiness. Other more technical descriptive attributes that require training would be unsuitable in a consumer test. Of course, JAR scales should not be used with trained panels as the judgment is about product likes and dislikes.

The endpoints must be true opposites. "Too thin" versus "Too thick" is a legitimate example. But "Too sour" versus "Too sweet" are not opposites even though they may show a negative correlation (as one goes up the other goes down) in a product. These should be separate scales. Avoid complex attributes like "creamy" which are made up of several combined qualities (smoothness, slipperyness, mouthcoating, viscosity, and dairy aroma can all contribute to creaminess). Avoid inherent negatives like bitterness unless they are actually desired in a product such as beer. Bitterness in milk has no optimum. Avoid vague positive terms like "natural." As with all scales in consumer tests, avoid redundancy. It does not make sense to have separate JAR scales for thin and thick when they are true opposites.

One must also be careful with the actions that are taken after obtaining JAR scale information. Any attempt to reduce or increase the intensity of an attribute may decrease the product's acceptance among those people who felt it was just-right. Also, the JAR ratings do not indicate how much to change the product to get a better result. Finally, foods and beverages are complex systems and any change in one attribute is likely to affect others. It is difficult to change sweetness without altering the sourness of a product due to taste mixture interactions, for example. Other problems and issues are discussed below.

14.5.3 Variations on Relative-to-Ideal Scaling

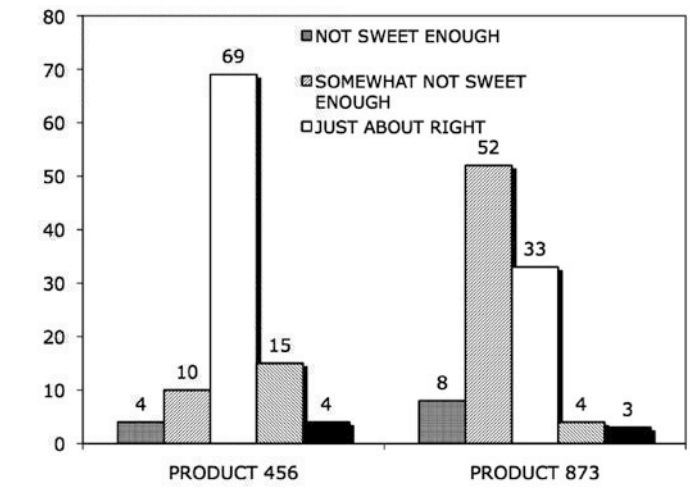
There are several types of scales that have used this idea of a central optimal point for the intensity of an attribute. A simple line scale was used by Johnson and Vickers (1987) and Vickers (1988) to study the optimization of sweetness. It was labeled "not nearly sweet enough" at the left end, "just-right" at the center, and "much too sweet" at the right. Pokorny and Davidek (1986) gave examples of several just-right scales to optimize the most important attributes of a product. In one case, the scale points were labeled to show how the product should be changed to be improved. At one end the response label was (the attribute should be) "very much stronger" and the opposite end was "very much weaker." The center point was "without change; it is optimal" and intermediate points were labeled "very slightly stronger" (or weaker), "slightly stronger" (or weaker), and "much stronger" (or weaker). Data were depicted graphically in terms of the percent of respondents giving the "optimal" response for each of nine key characteristics and an improvement in an appetizer product after reformulation was shown. Note that this phrasing is reversed from the common just-right scale where the descriptors refer to difference from ideal whereas in Pokorny and Davidek's example, the descriptors refer to the direction of change that would bring the product back toward the ideal level.

Another variation is to present a normal intensity scale for judging the product, and then ask respondents to place a second mark on the scale for their "ideal" product (van Trijp et al., 2007). An early example of this can be found in Szczesniak and Skinner (1975) in using a modified texture profile for consumers, where the ideal values for a whipped topping were also scored. This approach presents several advantages. First, the absolute intensity information is obtained as well as where the ideal product lies for that person on a scale. The second advantage is that the individual's scores can be expressed as deviation from ideal, so the just-right directional information can be obtained. Finally, and perhaps most importantly, an "ideal product profile" can be constructed if the data from the panelists are reasonably homogeneous, as well as mean deviations of each the test products from this ideal profile. The major limitation to this approach lies in the abilities of untrained consumers to act in such an analytical fashion, whether they can understand the terms that are being scaled and whether they can report their true feelings about an ideal product in the necessary detail.

A direct approach to product optimization is to have consumers adjust the level of some ingredient until they feel it is optimal (Pangborn and Pecore, 1982). This technique is commonly referred to as a "method of adjustment." However, this procedure must be performed in both directions, concentrating and diluting, to avoid context and/or adaptation effects which cause a consumer to stop too soon in the series (Mattes and Lawless, 1985). Hernandez and Lawless (1999) adapted the method of adjustment for liquid and solid food systems by sequentially weighing the amount added, then subtracting the amount consumed before the next addition in order to estimate the given concentration on any trial.

14.5.4 Analysis of JAR Data

When should product modifications be made based on JAR data? JAR data are actionable when there is an insufficient proportion of responses in the just-right category, when the data are asymmetric, or when there is evidence of a bimodal distribution. When should the current product be deemed acceptable at the current level? Obviously, a desirable set of responses on the just-right scale is centered on the optimum, is symmetric, and has low frequencies in the extremes of the scale. So the first step in analysis is to examine the frequencies across the scale, usually by plotting the data



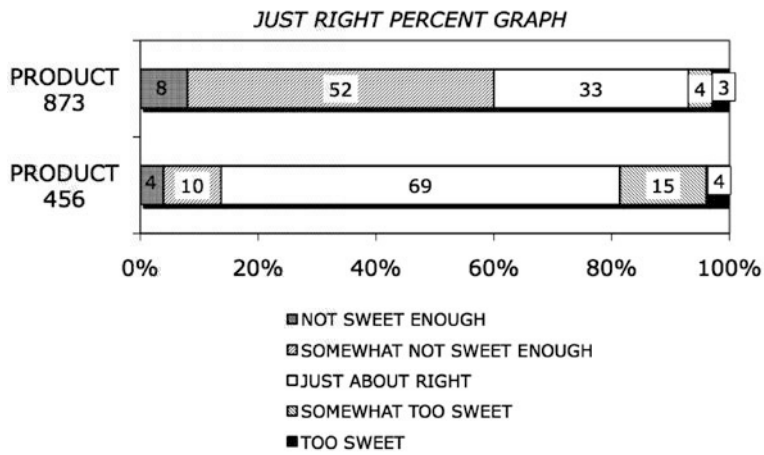


Fig. 14.8 Graphing JAR data. The *upper panel* shows a simple histogram in which product 456 is centered on the JAR category and symmetric, suggesting no change needs to be made. Product 873 is not symmetric (skewed) and has a large segment of "somewhat not sweet enough" suggesting an increase in the sweetener level could be an improvement. The *lower panel* shows the same data graphed as a constant sum bar graph.

as a simple bar graph (histogram) or a fractionated bar as shown in Fig. 14.8. One could easily be misled by examining only the means. For example, there might be two segments of consumers, one group that prefers a stronger level of the attribute and another group that prefers less. Examining only the mean from such opinions could lead to a false impression that the product was at or near the optimum level. So plot your data. The second question is, "Do I have enough just-right votes to leave the product as-is?" A common benchmark is something like 80% JAR votes in the center category (Rothman and Parker, 2009). One quantitative test for being skewed away from the just-right point is a simple one-sample *t*-test of the product's mean versus the center point's value on the scale. A simple nonparametric test is to compare the frequencies of those who are above JAR to those below JAR, while ignoring the JAR votes (Stone and Sidel, 2004). This can be a simple binomial probability test against an expected value of 0.5 (equal proportions). Both the *t*-test for the mean and the binomial test can provide evidence that the data are skewed higher or lower than the midpoint, for a single product.

The next consideration is how to compare multiple products to see if any are different or closer to the ideal. If each consumer has evaluated all the products (a within-subjects design or complete block), the chi-square statistic based on a cross-tabulation is not appropriate as it is based on the assumption that the data are from independent samples. Since the evaluations are related, several alternatives are available. Fritz (2009) provides a good discussion of these methods

and worked examples. For more than two products, the Cochran-Mantel-Haenszel ("CMH") method can be used, but it is computationally intensive and requires statistical software. For comparing any two products, the data can be cross-tabulated by the frequencies in a 3×3 table, after collapsing the data into three categories (above just-right, just-right, below just-right) for each product. This is shown in Fig. 14.9. Then the appropriate statistic is the Stuart–Maxwell test, as shown in the figure and Appendix B (Best and Raynor, 2001; Fleiss, 1981). If the data can be collapsed further, the simple McNemar test can be applied. For example, one might suspect that it is the "too sweet" category that is different for the two products. Then

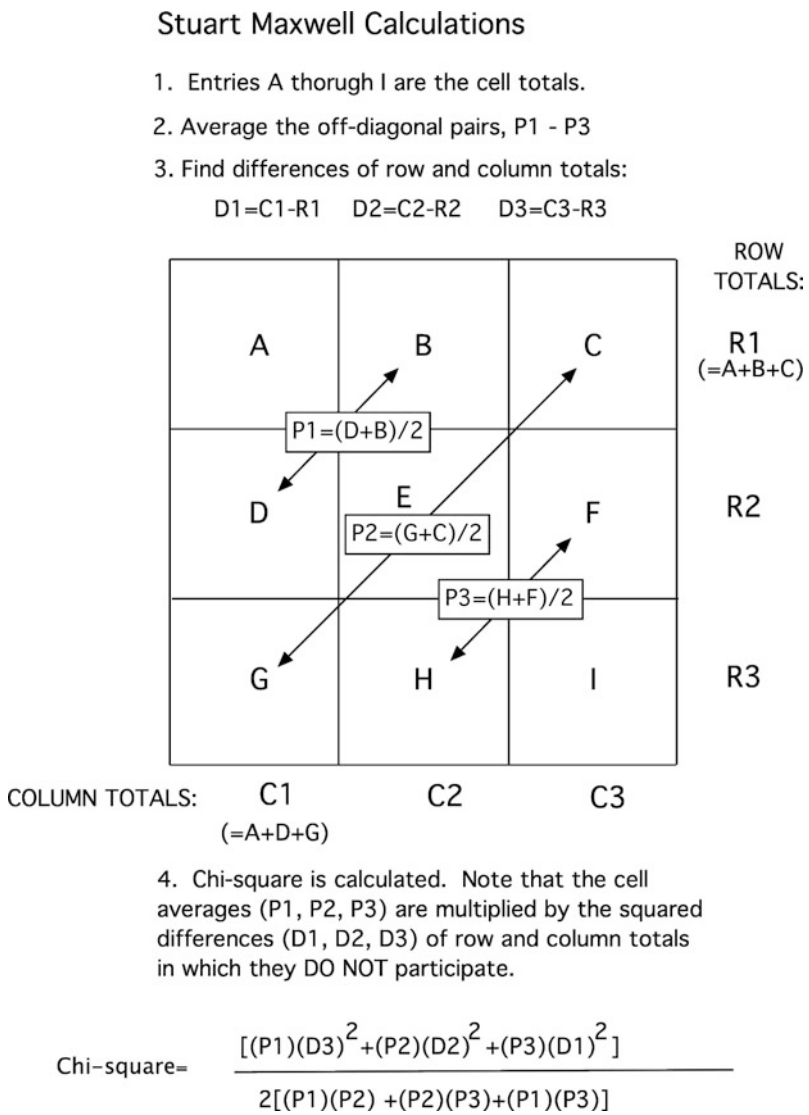


Fig. 14.9 The Stuart–Maxwell test for the differences between two products assessed on JAR scales. The data are first collapsed into three categories, those above just-right (i.e., too strong), those below just-right (i.e., too weak), and those falling in the just-right or just-about-right category. Entries A through I are the frequencies falling into each category considering their ratings for both products. The χ^2 test is done against a critical value for 2 df, which is 5.99. This value must be exceeded for a significant difference to be obtained. After a finding of a significant difference, the rows and columns can be collapsed into a 2×2 table for further analysis by the McNemar test to see if there are particular cells driving the Stuart–Maxwell result. See Appendix B for a worked example.

the McNemar can be applied to a 2×2 table pitting "too sweet" against the combined frequencies of the other categories. If the scales are continuous like a line scale or have contain seven or nine categories, then one can treat them as any other scaled response, with parametric statistics like *t*-tests and analysis of variance to compare means. If there are only a few response categories the data should be treated as categorical or ordinal.

14.5.5 Penalty Analysis or "Mean Drop"

Another source of valuable information can be from overall acceptance ratings collected in the same questionnaire. The JAR data can be combined with this information to assess the potential impact of being non-JAR (off from just-right) on the overall acceptability of the product. The approach is simple and proceeds as follows (from Schraidt, 2009):

(1) Separate the data into groups that were above, below, and at the just-right category.
(2) Calculate the mean hedonic scores from the acceptability scales for each of the three groups.
(3) Subtract the mean of the above-JAR group from the JAR group and likewise subtract the mean of the below-JAR group from the JAR group. (Note: it is important to use the JAR group's mean and not the overall data mean for this purpose.)
(4) Plot the resulting difference scores, the "mean drop," or penalty in a scatter plot of the mean drop versus the percentage of the total consumer panel in each category.

In this plot, a point that shows a large mean drop and a large percentage is a cause for concern and suggests that the product be modified in the appropriate direction. An example is shown in Table 14.2 and plotted in Fig. 14.10. In this plot we see that there is a large group who felt the product was too sweet and had a large mean drop or reduction in the overall acceptance score. The product development team might want to increase the sweetness level in a new version. There were two large groups, one of which felt the product was too thick and one too thin, but their penalties were small so no action need be taken. Regarding the fruit flavor strength, there was a small but strongly dissatisfied group in the upper left corner. Further research might be warranted to see whether a different style of product with a milder flavor would appeal to this

Table 14.2 Data for penalty analysis and mean drop in Fig. 14.10

	Mean	Drop	%
Attribute Sweetness			
Too sweet	6.2	–1.4	15
Just-about-right	7.6		50
Not sweet enough	5.2	–2.4	35
Thickness			
Too thick	6.5	–0.7	28
Just-about-right	7.2		37
Too thin	6.2	–1.0	35
Fruit flavor			
Too strong	5.2	–2.6	17
Just-about-right	7.8		63
Too weak	6.5	–1.3	20

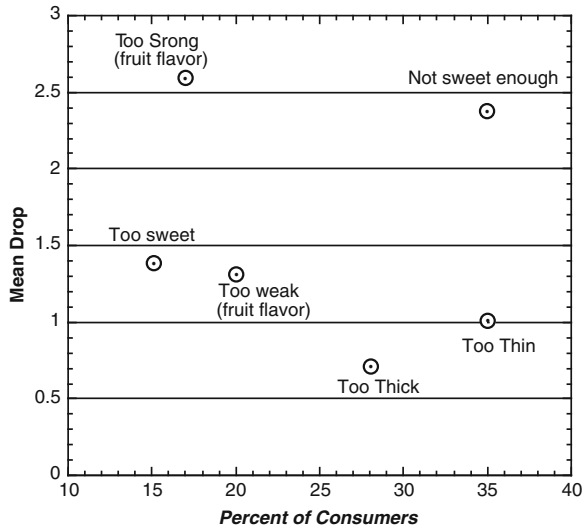


Fig. 14.10 Penalty analysis and mean drop for one hypothetical product. For each of the three scales shown in Table 14.2, consumers are categorized as rating each scale above, below or in the just-right category. The mean hedonic score (9-point scale value) is then calculated for each group. For groups above or below the JAR group, their scores are subtracted from the mean of the JAR group. This produces a "mean drop" score for each group. The percent of the consumer panel that fell into each category are plotted against their mean drops. Note that the category "not sweet enough" has both a high percentage and high mean drop, suggesting a potential improvement with increased sweetness.

group, perhaps leading to a lighter styled version for a consumer segment.

Various statistical tests and methods for modeling the penalty analysis are given in Rothman and Parker (2009). A simple approach is to make a 2×2 classification table of the above/below JAR versus likers/dislikers for the product and perform a chi-square test (Templeton, 2009). A significant result would suggest that one kind of JAR response was more detrimental to the product than the other. A parametric comparison could compare the group of above-JAR consumers to the below-JAR consumers to see whether their acceptance scores were in fact different (i.e., an independent groups *t*-test).

14.5.6 Other Problems and Issues with JAR Scales

In spite of their apparent simplicity, JAR scales are subject to a number of complications. These issues are discussed in Rothman and Parker (2009). Bear in mind that JAR scales only work with attributes where there is an optimum. If an attribute is a case of "more is always better" or "any of this is bad" then JAR scales are not appropriate. As noted before, one potential problem is that consumers may misinterpret the attribute. An example is the confusion that many people have concerning sour and bitter. A second concern is that consumers are generally integrative when they perceive a product as opposed to being analytical, and the JAR scale asks them to attend to a specific attribute in isolation. Consumers often show halo effects, in which one salient attribute can affect the ratings of other, logically unrelated attributes. Thus is it possible that a product could be rated less than ideal on sweetness, merely because the consumer is annoyed by some other taste or flavor. They can also bring in cognitive biases that have nothing to do with their own likes and dislikes. For example, if a person believes that salt is bad for you, the product could be rated as too salty, even if it is ideal. Some attributes are linked or may show trade-offs. Increasing sweetness may decrease sourness but the sweet/sour balance may be key as it is in some wines and fruit products. Some attributes may change with time or the hedonic reaction may change with time. Sweetness may seem acceptable at first, but may be cloying after consuming the entire portion. Portion size effects may influence JAR ratings. What seems good in a small bite may not seem so appealing in a larger portion. Other problems can occur when JAR scales are used in isolation without asking any additional intensity-related questions. Thus two groups of respondents might both mark "just-right" but one might think the product is very strong (the level they prefer) while the other group thinks the product is fairly mild (the level they prefer). So the results might mislead product developers into thinking that there is a homogeneous population while there are really two consumer segments. The sensory professional should consider the usefulness of collecting both intensity and JAR information on the consumer questionnaire.

The centering tendency or bias is particularly problematic for JAR scales, especially in a multi-product test. The centering bias in this case is a tendency to put the product which is intermediate in intensity (i.e., the middle product) at the just-right point. This could lead to a false conclusion that the middle product was just-right when in fact the true optimum could still be higher or lower. Johnson and Vickers (1988) compared several methods for dealing with the centering problem, based on approaches of McBride (1982) and Poulton (1989). Both of these methods involve testing multiple ranges of products to interpolate the JAR point based on different ranges of the key ingredient. They are discussed further in Chapter 9. It may not always be possible to do the multiple sessions required to interpolate the true optimum, so the sensory scientist should be aware that a false reading for the middle product in a series can be a problem with this method.

14.6 Behavioral and Context-Related Approaches

It is difficult to measure a stable attitude toward a food when it is tasted in isolation. Our likes and dislikes may certainly change as a function of the context of a meal, the time of day, or the number of times we have consumed the food recently. Foods and beverages may be liked more or less depending upon the temperature at which they are served (Kahkonen et al., 1995) and whether that temperature is consistent with consumer's expectations (Cardello and Maller, 1982a). A person's recent history with similar foods can be an important influence. Some people are prone to seek a

high degree of variety in their diet (van Trijp et al., 1992) and may tire of eating foods with consistent or similar sensory properties (Vickers, 1988). Eating steak or lobster may seem highly appealing in the abstract, but consuming steak or lobster for ten days in a row will certainly change a person's feelings about the item toward the end of that period. Preferences are also specific to foods that are combined. So although we may like ketchup in general, ketchup on mashed potatoes may or may not seem appealing to a given individual. A person's historical preferences across many foods may fail to predict their hedonic acceptance scores for items in an actual tasting (Cardello and Maller, 1982b). Context and expectations can affect simple hedonic judgments (Deliza and MacFie, 1996; Meiselman, 1992; Schifferstein, 1995), so we should expect them to have important effects in actual food choice and consumption situations. The eating situation can have a big effect (Edwards et al., 2003; King et al., 2007). Habit, experience, context, and attitudes are important contributors to the actual consumption of a food in a specific situation. To address some of these limitations of simple acceptance testing some more behaviorally oriented approaches have been utilized.

14.6.1 Food Action Rating Scale (FACT)

An example of a behaviorally oriented approach to scaling food acceptability was devised by Schutz (1965). He developed a scale based on attitudes and actions, combining some statements about frequency of consumption ("I would eat this food every opportunity I had") and some motivationally related statements ("I would eat this food only if I were forced to") to produce a more action-oriented index of food acceptance. This was called the Food Action Rating Scale or FACT. The complete list of descriptors is shown in Table 14.3. Although Schutz reasoned that the behaviors might not always match up with acceptance as scaled on a traditional 9-point hedonic scale, a study of the correspondence of the two measures showed a high degree of positive correlation (r = +0.97 in a questionnaire study of food likes). Data from the FACT scale gave lower mean values but less skew compared to the 9-point hedonic scale. In spite of this correlation, the scales are not necessarily interchangeable.

Table 14.3 Descriptors in the food action rating scale (FACT)

I would eat this food every opportunity I had
I would eat this very often
I would frequently eat this
I like this and would eat it now and then
I would eat this if available but would not go out of my way
I do not like it but would eat it on an occasion
I would hardly ever eat this
I would eat this only if there were no other food choices
I would eat this only if I were forced to

After Schutz (1965)

14.6.2 Appropriateness Scales

In related work, Shutz carried the contextual theme one step farther. Appropriateness ratings can be used to assess the effects of context associated with hedonic responses to food (Schutz, 1988). For example, one may really like pizza but when asked to rate one's liking at 8 A.M. may not find the pizza appealing since this is an inappropriate time to consume pizza, at least for most people. So pleasantness on a purely sensory basis and appropriateness in a context may not always be completely parallel. Inappropriate contexts for a given culture may override influences of sensory liking (Lahteenmaki and Tuorila, 1997). While it is certainly possible to poll consumers about the appropriateness of foods in different contexts, the sensory scientist should be sensitive to the testing burden that can arise in asking too many questions. If 15 foods are evaluated for appropriateness in 20 contexts each consumer panelist would have to rate 300 scales.

Appropriateness judgments are traditionally done for both a food and a usage statement like "eaten for breakfast." The scale commonly used is a 7-point scale ranging from "never" to "always." Note that this is not a scale that ranges from "not appropriate" to "very appropriate." Rather, it attempts to tap into a notion of the frequency with which an item would apply for that usage situation, in a graded fashion. The questionnaire often takes the form of a matrix or grid, with foods as rows and usage statements as columns. A good review of appropriateness questions and sample applications can be found in Schutz (1994). In some studies comparing appropriateness to the 9-point hedonic scale, Cardello and Schutz (1996) showed that foods with equal acceptability could differ

dramatically in their appropriateness for various situations. Although this is perhaps not surprising, it should be noted that liking and appropriateness are often highly correlated. Foods that are highly liked find appropriate uses in a variety of situations. However, the work of Cardello and Schutz demonstrates that they are not equivalent. This study also noted that acceptability scores were unaffected by asking the appropriateness questions. So there seems to be no harm in asking for this additional information, keeping in mind the added burden on test subjects due to the length of the questionnaire.

Data from appropriateness judgments can be analyzed by principal component analysis to determine the underlying features common to the different foods and contexts. Jack et al. (1994) used a repertory grid triad method to derive all potential use occasions for cheeses. The appropriateness of the use of each of 16 cheeses was then evaluated on line scales anchored by suitable and unsuitable. They found that the melting characteristics and the texture of the cheeses were the major factors affecting the appropriateness of cheese for different occasions. In consumer products, it is often important to have fragrances that are consonant with the intended use of a product. The type of fragrance that is appropriate in a shampoo may not work well as a fragrance used to mask insecticide chemical smells and vice versa (Jellinek, 1975). So in screening candidate fragrances for specific product applications, it is necessary to go beyond simple hedonics and ask people about the fit of the smell to the intended product. A fruity-floral fragrance may be appealing as an air freshener, but may seem out of place in an institutional heavy-duty sanitizing cleaner. An appropriateness scale in this kind of situation would range from "not appropriate" to "very appropriate" in contrast to the frequency scale of Schutz.

14.6.3 Acceptor Set Size

Another variation in hedonic assessment brings the food acceptance measure back to a simple count of the proportion of people who find the product appealing. Norback and colleagues (Beausire et al., 1988; LaGrange and Norback, 1987) examined the proportion of acceptors as a variable in product optimization. The acceptor set size was defined as the proportion of consumers who find the product acceptable. LaGrange and Norback (1987) reasoned quite logically that the causal chain in optimizing the sensory appeal of a product should consider an acceptability measure, such as the proportion of acceptors, to be driven by a set of sensory attributes. The sensory attributes in turn would be determined by the physical characteristics of the food, ingredient variables, and so on. This is simply a straightforward psychophysical model extended to include multiple attributes and a hedonic or behavioral consequence. They reasoned that those variables that have a strong impact on changing the acceptor set size (i.e., produce a steep slope or high rate of change) would be most influential in optimization. This is much in the tradition of modeling acceptance as a function of a set of contributing attributes (e.g., Moskowitz and Krieger, 1995). Beausire et al. (1988) made use of this approach in a linear programming model to explore the relationship between acceptor set size, toughness scores, and ingredients in turkey bratwurst. Various ingredient combinations were explored for their effects on the toughness attribute and the resulting acceptability function.

This approach makes use of some minimal information, basically whether people find the product acceptable or not. In the Beausire et al. (1988) study, participants were simply asked to decide "yes" or "no" as to whether they found the product acceptable. On the positive side, this measure taps into a fundamental concern for food marketers, the size of the pool of potential users. On the negative side, the measure is dichotomous and lacks the graded richness of information available on the 9-point acceptance scale. It is also possible to think of this measure as a kind of preference split, but based on only one product versus the consumer's minimal expectations for the category. As Stone and Sidel (2004) pointed out, preference may not always match up with acceptance. People who have no liking at all for the category do not belong in the test, but they may express a preference if mistakenly included in the study. It is also possible to have a preference win for a product based on the proportion of people preferring one product over another. However, a minority with strong opinions in the opposite direction may give the preferred product low hedonic ratings with the result that the mean acceptability scores are reversed from the preference split. For example, product A was higher than B in acceptance, but product B was preferred over A in simple choice (see Stone

and Sidel, 2004, for numerical example). This same problem could potentially arise with the acceptor set. There might be fair proportion of people finding the product acceptable, but a group that dislikes it strongly since it is not in their favorite or accustomed style. Acceptor set size and simple preference/choice might miss this fact. Examination of acceptability scores, on the other hand, would uncover both the proportion and the degree of dislike in the minority segment.

14.6.4 Barter Scales

Another approach to food acceptance was developed for use in predicting meal combinations. Lawless (1994) examined meal combinations for military field rations. The dependent measure was based on the number of chocolate bars that a soldier would trade for individual items and for meal combinations. This measure was based on the observation that trading for food items within the prepackaged meals did occur in the field that the chocolate bar was a highly desirable item and that it could function as a kind of "common coinage" for swapping. The overall goal was to develop a measure that could predict the value or total appeal of meal combinations and then to use this measure along with nutritional information and other constraints in developing a linear optimization model for ration improvement. One concern of the product developers was that the 9-point scale was too constrained and would lack the additive properties necessary for an estimate of total hedonic value in meal combinations. For example, two items might be rated an eight and a five on the hedonic scale, but there is no scale point corresponding to 13 or the sum of the reactions to the two items since the scale is bounded by nine points. It was hypothesized that the barter value of the items and meal combinations would show more reasonable patterns of additivity that could be used in linear programming.

The values in chocolate bars of 33 individual items were estimated and then they were combined to form two-item combinations and five-item meals, similar or identical to actual field package combinations. The question of interest was whether the values of the meals and combinations could be predicted by the sum or some simple linear combination of the values of the items. The data showed almost perfect additivity with one notable exception. The data showed an "a la carte" effect. That is, the value expected in trade for the meal was one chocolate bar less than the sum of the individual items. This is similar to the common pricing of meals versus a la carte or individual items where the sum of the individual prices would be higher than the price paid for the actual meal as a combination. The utility of this method may depend upon a consistent positive value of the bartered item among the participants.

14.7 Conclusions

The assessment of blind-labeled product acceptability is one of the cornerstones of sensory evaluation. Consumer acceptance is essential "bottom line" information for product developers and marketers alike. A variety of useful methods are available to the sensory scientist in order to assess the appeal of products and the relative preferences among a set of choices. Choice itself is fundamental to consumer behavior, as it is the decision process that we all make when faced with a number of different foods for purchase or for use in a meal. Acceptance can be related to other properties of foods such as the descriptive profile of a product or to physical ingredient, processing and packaging variables. We can study how the appeal of a product declines over periods of storage, in the distribution chain and over its shelf life.

In spite of the obvious importance of product acceptability, this group of sensory methodologies is prone to misuse, misinterpretation, and challenges from other fields. In particular, marketing researchers may not understand the importance of testing products on a blind-labeled basis, i.e., with a minimal conceptual framework. The rule of thumb for sensory research is to present only enough information so that the product is evaluated within the correct frame of reference, usually the product category. So participants in a sensory acceptance test are given a product with a random three-digit code, they know only that it is a test of scrambled eggs, and they are unaware that the eggs are frozen, reconstituted, microwaved, cholesterol reduced, or any number of other factors that might eventually form part of the product concept. The sole question is whether they are appealing to scrambled egg consumers on the basis of their sensory attributes (taste, texture, appearance, etc.).

A common complaint among marketers is that the blind test is unrealistic since the product will not appear on the store shelves in its unlabeled form without package and concept (see Garber et al., 2003, for a critique of sensory practice from a marketing perspective). However, they miss the point. The sensory test is the only way to assess true sensory appeal without the biasing effects of conceptual labeling (Gacula et al., 1986) and as such it gives essential feedback to product developers as to whether they have truly met their target profiles. Without the blind test, no one can tell. The product may succeed or fail in the concept-laden market research test for any number of reasons. Given the tendency of consumers to integrate their information, to show halo effects and other biases, you cannot always trust their stated reasons for liking in a questionnaire from a concept test. As stated above, high-sensory acceptability does not insure that the product will be a marketplace success. Purchase probability (and more importantly, repurchase) depends upon price, concept, positioning, promotions, advertising, package information, consumer awareness, nutritional characteristics, and many other factors (Garber et al., 2003). However, the sensory appeal is the essential "platform" without which the product is unlikely to succeed. This platform—of sensory-based acceptance—provides the foundation for successful marketers to then apply their artistry to sell the product to consumers in the real world.

References

Bahn, K. D. 1989. Cognitive and perceptually based judgments in children's brand discriminations and preferences. Journal of Business and Psychology, 4, 183–197.

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

Beausire, R. L. W., Norback, J. P. and Maurer, A. J. 1988. Development of an acceptability constraint for a linear programming model in food formulation. Journal of Sensory Studies, 3, 137–149.

Best, D. J. and Rayner, J. C. W. 2001. Application of the Stuart test to sensory evaluation data. Food Quality and Preference, 12, 353–357.

Birch, L. L. 1979. Dimensions of preschool children's food preferences. Journal of Nutrition Education, 11, 77–80.

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

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

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

Booth, D. A. 1994. Flavour quality as cognitive psychology: The applied science of mental mechanisms relating flavour descriptions to chemical and physical stimulation patterns. Food Quality and Preference, 5, 41–54.

Booth, D. A. 1995. The cognitive basis of quality. Food Quality and Preference, 6, 201–207.

Bower, J. A. and Boyd, R. 2002. Effect of health concern and consumption patterns on measures of sweetness by hedonic and just right scales. Journal of Sensory Studies, 18, 235–248.

Byer, A. J. and Saletan, L. T. 1961. A new approach to flavor evaluation of beer. Wallerstein Laboratory Communications, 24, 289–300.

Cardello, A. V. and Maller, O. 1982a. Acceptability of water, selected beverages and foods as a function of serving temperature. Journal of Food Science, 47, 1549–1552.

Cardello, A. V. and Maller, O. 1982b. Relationships between food preferences and food acceptance ratings. Journal of Food Science, 47, 1553–1557, 1561.

Cardello, A. V. and Schutz, H. G. 1996. Food appropriateness measures as and adjunct to consumer preference/acceptability evaluation. Food Quality and Preference 7, 239–249.

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

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

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

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

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

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

Chung, S.-J., and Vickers, A. 2007b. Influence of sweetness on the sensory-specific satiety and long-term acceptability of tea. Food Quality and Preference, 18, 256–264.

Coetzee, H. 1996. The successful use of adapted paired preference, rating and hedonic methods for the evaluation of

acceptability of maize meal produced in Malawi. Abstract, 3rd Sensometrics Meeting, June 19–21, 1996, Nantes, France, pp. 35.1–35.3.

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

Deliza, R. and MacFie, H. J. H. 1996. The generation of sensory expectation by external cues and its effect on sensory perception and hedonic ratings: A review. Journal of Sensory Studies, 11, 103–128.

Edwards, J. A., Meiselman, H. L., Edwards, A. and Lesher, L. 2003. The influence of eating location on the acceptability of identically prepared foods. Food Quality and Preference, 14, 647–652.

El Dine, A. N. and Olabi, A. 2009. Effect of reference foods in repeated acceptability tests: Testing familiar and novel foods using 2 acceptability scales. Journal of Food Science, 74, S97–S106.

Engen, T. 1974. Method and theory in the study of odor preferences. In: A. Turk, J. W. Johnson, Jr. and D. G. Moulton (eds.), Human Responses to Environmental Odors. Academic, New York, pp. 121–141.

Engen, T. 1978. The origin of preferences in taste and smell. In: J. H. A. Kroeze (ed.), Preference Behaviour and Chemoreception. Information Retrieval, London, pp. 263–273.

Engen, T., Lipsitt, L. P. and Peck, M. 1974. Ability of newborn infants to discriminate sapid substances. Developmental Psychology, 10, 741–744.

Epler, S., Chambers, E. and Kemp, K. E. 1997. Hedonic scales are better predictors than just right scales of optimal sweetness in lemonade. Journal of Sensory Studies, 13, 191–197.

Fleiss, J. L. 1981. Statistical Methods for Rates and Proportions, Second Edition. Wiley, New York.

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

Fritz, C. 2009. Appendix G: Methods for determining whether JAR distributions are similar among products (Chi-square, Cochran-Mantel-Haenszel (VMH), Stuart-Maxwell, McNemar). In: L. Rothman and M. J. Parker (eds.), Just-About-Right Scales: Design, Usage, Benefits, and Risks. ASTM Manual MNL63, ASTM International, Conshohocken, PA, pp. 29–37.

Gacula, M. C., Rutenbeck, S. K., Campbell, J. F., Giovanni, M. E., Gardze, C. A. and Washam, R. W. 1986. Some sources of bias in consumer testing. Journal of Sensory Studies, 1, 175–182.

Garber, L. L., Hyatt, E. M. and Starr, R. G. 2003. Measuring consumer response to food products. Food Quality and Preference, 13, 3–16.

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

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

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

Greenhoff, K. and MacFie, H. J. H. 1994. Preference mapping in practice. In: H. J. H. MacFie and D. M. H. Thomson (eds.), Measurement of Food Preferences. Blackie Academics, London, pp. 137–166.

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

Head, M. K., Giesbrecht, F. G. and Johnson, G. N. 1977. Food acceptability research: Comparative utility of three types of data from school children. Journal of Food Science, 42, 246–251.

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

Helgensen, H., Solheim, R. and Naes, T. 1997. Consumer preference mapping of dry fermented lamb sausages. Food Quality and Preference, 8, 97–109.

Hernandez, S. V. and Lawless, H. T. 1999. A method of adjustment for preferred levels of capsaicin in liquid and solid food systems among panelists of two ethnic groups and comparison to hedonic scaling. Food Quality and Preference, 10, 41–49.

Hottenstein, A. W., Taylor, R., and Carr, B. T. 2008. Preference segments: A deeper understanding of consumer acceptance or a serving order effect? Food Quality and Preference, 19, 711–718.

Hough, G., Bratchell, N. and Wakeling, I. 1992. Consumer preference of Dulce de Leche among students in the United Kingdom. Journal of Sensory Studies, 7, 119–132.

Jack, F. R., Piggott, J. R. and Paterson, A. 1994. Use and appropriateness in cheese choice, and an evaluation of attributes influencing appropriateness. Food Quality and Preference, 5, 281–290.

Jellinek, J. S. 1975. The Use of Fragrance in Consumer Products. Wiley, New York.

Johnson, J. and Vickers, Z. 1987. Avoiding the centering bias or range effect when determining an optimum level of sweetness in lemonade. Journal of Sensory Studies, 2, 283–292.

Johnson, J. R. and Vickers, Z. 1988. A hedonic price index for chocolate chip cookies. In: D. M. H. Thomson (ed.), Food Acceptability. Elsevier Applied Science, London, pp. 135–141.

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

Kahkonen, P., Tuorila, H. and Hyvonen, L. 1995. Dairy fact content and serving temperature as determinants of sensory and hedonic characteristics of cheese soup. Food Quality and Preference, 6, 127–133.

Keskitalo, K., Knaapila, A., Kallela, M., Palotie, A., Wessman, M., Sammalisto, S., Peltonen, L., Tuorila, H. and Perola, M. 2007. American Journal of Clinical Nutrition, 86, 55–63.

Kim, K.-O. and O'Mahony, M. 1998. A new approach to category scales of intensity I: Traditional versus rank-rating. Journal of Sensory Studies, 13, 241–249.

Kimmel, S. A., Sigman-Grant, M. and Guinard, J.-X. 1994. Sensory testing with young children. Food Technology, 48(3), 92–94, 96–99.

King, S. C., Meiselman, H. L., Hottenstein, A. W., Work, T. M. and Cronk, V. 2007. The effects of contextual variables on food acceptability: A confirmatory study. Food Quality and Preference, 18, 58–65.

Koster, E. P., Couronne, T. Leon, F. Levy, C. and Marcelino, A. S. 2003. Repeatability in hedonic sensory measurement: A conceptual exploration. Food Quality and Preference, 14, 165–176.

Kroll, B. J. 1990. Evaluating rating scales for sensory testing with children. Food Technology, 44(11), 78–80, 82, 84, 86.

Lagrange, V. and Norback, J. P. 1987. Product optimization and the acceptor set size. Journal of Sensory Studies, 2, 119–136.

Lahteenmaki, L. and Tuorila, H. 1997. Item by use appropriateness of drinks varying in sweetener and fat content. Food Quality and Preference, 8, 85–90.

Lawless, H. T. 1977. The pleasantness of mixtures in taste and olfaction. Sensory Processes, 1, 227–237.

Lawless, H. T. 1994. Contextual and measurement aspects of acceptability. Final Report #TCN 94178, U. S. Army Research Office.

Lawless, H. T. 2010. Commentary on "Comparative performance of the nine-point hedonic hybrid and self-adjusting scales in generation of internal preference maps." Food Quality and Preference, 21, 165–166.

Lawless, H. T. and Malone, G. J. 1986. A comparison of scaling methods: Sensitivity, replicates and relative measurement. Journal of Sensory Studies, 1, 155–174.

Lawless, H. T., Hammer, L. D. and Corina, M. D. 1982–1983. Aversions to bitterness and accidental poisonings among preschool children. Journal of Toxicology: Clinical Toxicology, 19, 951–964.

Lawless, H. T., Popper, R. and Kroll, B. 2010. Comparison of the labeled affective magnitude (LAM) scale, an 11-point category scale and the Traditional nine-point hedonic scale. Food Quality and Preference, 21, 4–12.

Lavanaka, N. and Kamen, J. 1994. Magnitude estimation of food acceptance. Journal of Food Science, 59, 1322–1324.

Mattes, R. D. and Lawless, H. T. 1985. An adjustment error in optimization of taste intensity. Appetite, 6, 103–114.

McBride, R. 1990. The Bliss Point Factor. Macmillan, South Melbourne, NSW (Australia).

McBride, R. L. 1982. Range bias in sensory evaluation. Journal of Food Technology, 17, 405–410.

McDaniel, M. R. and Sawyer, F. M. 1981. Preference testing and sensory evaluation: Magnitude estimation vs. the 9-point hedonic scale. Journal of Food Science, 46, 182–185.

McEwan, J. 1996. Preference mapping for product optimization. In: Multivariate Analysis of Data in Sensory Science. Elsevier Applied Science, London, pp. 71–102.

Mead, R. and Gay, C. 1995. Sequential design of sensory trials. Food Quality and Preference, 6, 271–280.

Meiselman, H. L. 1992. Methodology and theory in human eating research. Appetite, 19, 49–55.

Meilgaard, M., Civille, G. V. and Carr, B. T. 1991. Sensory Evaluation Techniques, Second Edition. CRC, Boca Raton.

Moskowitz, H. R. 1980. Psychometric evaluation of food preferences. Journal of Foodservice Systems, 1, 149–167.

Moskowitz, H. R. 1986. New Directions for Product Testing and Sensory Analysis of Foods. Food and Nutrition, Westport, CT.

Moskowitz, H. R. and Krieger, B. 1995. The contribution of sensory liking to overall liking: An analysis of six food categories. Food Quality and Preference, 6, 83–90.

O'Mahony, M., Park, H., Park, J. Y. and Kim, K.-O. 2004. Comparison of the statistical analysis of hedonic data using analysis of variance and multiple comparisons versus and R-index analysis of the ranked data. Journal of Sensory Studies, 19, 519–529.

Pagliarini, E., Gabbiadini, N. and Ratti, S. 2005. Consumer testing with children on food combinations for school lunch. Food Quality and Preference, 16, 131–138.

Pagliarini, E., Ratti, S., Balzaretti, C. and Dragoni, I. 2003. Evaluation of a hedonic method for measuring the acceptability of school lunches by children. Italian Journal of Food Science, 15, 215–224.

Pangborn, R. M. and Pecore, S. D. 1982. Taste perception of sodium chloride in relation to dietary intake of salt. American Journal of Clinical Nutrition, 35, 510–520.

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

Peryam, D. R. 1989. Reflections. In: Sensory Evaluation. Celebration of our Beginnings. ASTM, Committee E-18 on Sensory Evaluation of Materials and Products, Philadelphia, pp. 21–30.

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

Peryam, D. R. and Pilgrim, F. J. 1957. Hedonic scale method of measuring food preferences. Food Technology, September 1957, 9–14.

Pokorny, J. and Davidek, J. 1986. Application of hedonic sensory profiles for the characterization of food quality. Die Nahrung, 8, 757–763.

Popper, R. and Kroll, B. R. 2005. Just-about-right scales in consumer research. Chemo Sense, 7, 1–6.

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

Resurreccion, A. V. A. 1998. Consumer Sensory Testing for Product Development. Aspen, Gaithersburg, MD.

Rohm, H. and Raaber, S. 1991. Hedonic spreadability optima of selected edible fats. Journal of Sensory Studies, 6, 81–88.

Rothman, L. and Parker, M. J. 2009. Just-About-Right Scales: Design, Usage, Benefits, and Risks. ASTM Manual MNL63, ASTM International, Conshohocken, PA.

Schifferstein, H. J. N. 1995. Contextual shifts in hedonic judgment. Journal of Sensory Studies, 10, 381–392.

Schmidt, H. J. and Beauchamp, G. K. 1988. Adult-like odor preference and aversions in three-year-old children. Child Development, 59, 1136–1143.

Schraidt, M. F. 1991. Testing with children: Getting reliable information from kids. ASTM Standardization News, March 1991, 42–45.

Schraidt, M. 2009. Appendix L: Penalty analysis or mean drop analysis. In: L. Rothman and M. J. Parker (Eds.), Just-About-Right Scales: Design, Usage, Benefits, and Risks. ASTM Manual MNL63, ASTM International, Conshohocken, PA, pp. 50–47.

Schutz, H. G. 1965. A food action rating scale for measuring food acceptance. Journal of Food Science, 30, 365–374.

Schutz, H. G. 1988. Beyond preference: Appropriateness as a measure of contextual acceptance. In: D. M. H. Thomson (ed.), Food Acceptability. Elsevier, London, pp. 115–134.

Schutz, H. G. 1994. Appropriateness as a measure of the cognitive-contextual aspects of food acceptance. In: H. J. H. MacFie and D. M. H. Thomson (eds.), Measurement of Food Preferences. Chapman and Hall, pp. 25–50.

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

Shepherd, R., Smith, K., and Farleigh, C. A. 1989. The relationship between intensity, hedonic and relative-to-ideal ratings. Food Quality and Preference 1, 75–80.

Stone, H. and Sidel, J. 2004. Sensory Evaluation Practices, Third Edition. Elsevier Academic, San Diego.

Szczesniak, A. S. and Skinner, E. Z. 1975. Consumer texture profile method. Journal of Food Science, 40, 1253–1256.

Templeton, L. 2009. Appendix R: Chi-square. In: L. Rothman and M. J. Parker (eds.), Just-About-Right Scales: Design, Usage, Benefits, and Risks. ASTM Manual MNL63, ASTM International, Conshohocken, PA, pp. 75–81.

Tepper, B. J., Shaffer, S. E. and Shearer, C. M. 1994. Sensory perception of fat in common foods using two scaling methods. Food Quality and Preference, 5, 245–252.

van Trijp, H. C. M., Lahtennmaki, L. and Tuorila, H. 1992. Variety seeking in the consumption of spread and cheese. Appetite, 18, 155–164.

van Trijp, H. C. M., Punter, P. H., Mickartz, F. and Kruithof, L. 2007. The quest for the ideal product: Comparing different methods and approaches. Food Quality and Preference, 18, 729–740.

Vickers, A. 1988. Sensory specific satiety in lemonade using a just right scale for sweetness. Journal of Sensory Studies, 3, 1–8.

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

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

Wakeling, I. N. and MacFie, H. J. H. 1995. Designing consumer trials balanced for first and higher orders of carry-over effect when only a subset of k samples from t may be tested. Food Quality and Preference, 6, 299–308.

Wright, A. O. 2007. Comparison of Hedonic, LAM, and other scaling methods to determine Warfighter visual liking of MRE packaging labels, includes web-based challenges, experiences and data. Presentation at the 7th Pangborn Sensory Science Symposium, Minneapolis, MN, 8/12/07. Supplement to Abstract Book/Delegate Manual.

Yao, E., Lim, J., Tamaki, K., Ishii, R., Kim, K.-O. and 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.