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

14.2 Hedonic Scaling: Quantification of Acceptability

14.3 Recommended Procedure

14.4 Other Acceptance Scales

14.5 Just-About-Right Scales

14.6 Behavioral and Context-Related Approaches

14.7 Conclusions

References

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
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).

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,
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

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 nonsensory 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).
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
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 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).](image)
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