

The 9-point hedonic scale and hedonic ranking in food science: some reappraisals and alternatives

Sukanya Wichchukit^{a,b} and Michael O'Mahony^{c*}

Abstract

The 9-point hedonic scale has been used routinely in food science, the same way for 60 years. Now, with advances in technology, data from the scale are being used for more and more complex programs for statistical analysis and modeling. Accordingly, it is worth reconsidering the presentation protocols and the analyses associated with the scale, as well as some alternatives. How the brain generates numbers and the types of numbers it generates has relevance for the choice of measurement protocols. There are alternatives to the generally used serial monadic protocol, which can be more suitable. Traditionally, the 'words' on the 9-point hedonic scale are reassigned as 'numbers', while other '9-point hedonic scales' are purely numerical; the two are not interchangeable. Parametric statistical analysis of scaling data is examined critically and alternatives discussed. The potential of a promising alternative to scaling itself, simple ranking with a hedonic R-Index signal detection analysis, is explored in comparison with the 9-point hedonic scale.

© 2014 Society of Chemical Industry

Keywords: 9-point hedonic scale; cognitive strategies; data analysis; alternative protocols; ranking; hedonic R-Index

INTRODUCTION

Part of food product development and the launching of new products in the market require some measure of whether the products are liked or not by the appropriate consumers. There have been many rating scales developed for measuring degree of liking^{1,2} of which the Labeled Hedonic Scale, sometimes called the LIM scale,^{3,4} and the LAM scale^{5,6} are more recent developments. The latter has been reviewed.⁷

However, in food science, probably the most used scale over the last 60 years has been the 9-point hedonic scale^{8,9} introduced as an aid to menu planning for US soldiers in their canteens. The scale comprises a series of nine verbal categories ranging from 'dislike extremely' to 'like extremely' and is described as such in various sensory texts (e.g.^{10–12}). For subsequent quantitative and statistical analysis, the verbal categories are generally assigned numerical values, ranging from 'like extremely' as '9' to 'dislike extremely' as '1'.^{13,14} Here, a scale like the traditional 9-point hedonic scale which is comprised of a series of labels, will be called a 'words only' scale. A hedonic scale which is purely numerical and which may only have labels at each end and sometimes in the middle, will be called a 'numbers only' scale (see Fig. 1).

This paper will not be a general review of hedonic scales, documenting their form and their various applications. This has already been done recently.¹ Instead, the authors examine the traditional protocols and analyses used with the applications of the 9-point hedonic scale to consumers, while suggesting some alternatives. Furthermore, an alternative to scaling itself – simple ranking – will be considered. With a signal detection R-Index analysis, this latter method provides the same information as is obtained using mean values from a numerical hedonic scale, without requiring consumers ever to use a scale.

LIKING AND PREFERENCE

For an appraisal of the 9-point hedonic scale, it is best to consider why it is being used. What are the goals of the measurement? The original 'words only' 9-point hedonic scale is a scale of liking. Consumers are required to assess a product and report how much they like it. As an aid to planning menus and for similar tasks it is useful; foods that are liked by many customers can remain on the menu while those that are disliked by many customers can be removed. The goal here is not to compare the comparative degree of liking between foods but merely to register whether a food is liked well enough to remain on the menu. The judgements are more absolute than comparative.

On the other hand, in food science, the scale is generally used comparatively. Logically, it can be inferred from this 'words only' scale that if food 'A' is 'liked extremely' and food 'B' is 'liked very much' or 'liked moderately', then food 'A' is liked more or is preferred to food 'B'. Used in this way, the scale becomes one of preference. Thus, assigning numbers 1–9 to the verbal responses on the 'words only' hedonic scale would be assigning at least an ordinal measure of preference to the products in question.¹³

* Correspondence to: Michael O'Mahony, Dept. Food Science & Technology, University of California, Davis, CA 95616, USA. E-mail: maomahony@ucdavis.edu

a Department of Food Engineering, Kasetsart University, Kamphaeng Saen, Nakorn-pathom, Thailand

b Center of Excellence in Agricultural and Food Machinery, Kasetsart University, Thailand

c Department of Food Science & Technology, University of California, Davis, California, USA

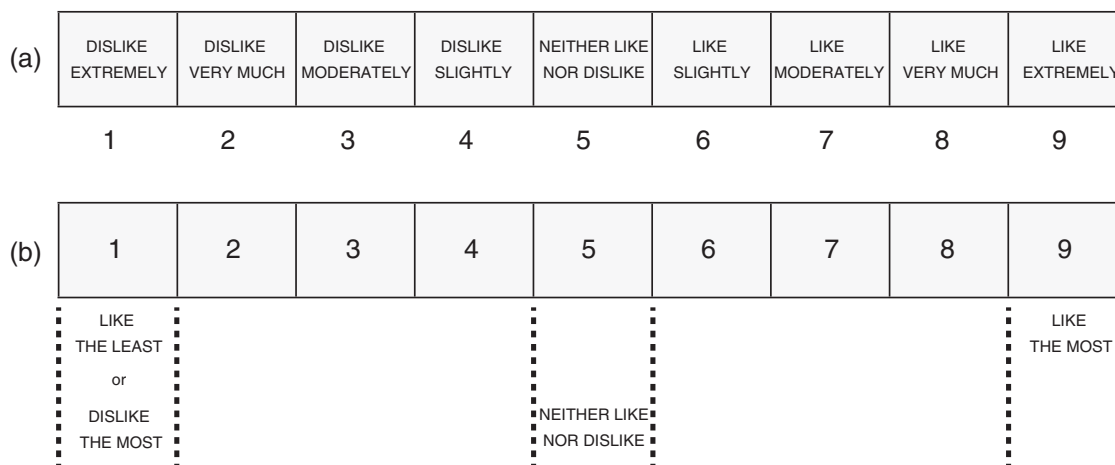


Figure 1. Versions of the 9-point hedonic scale. The traditional ‘words only’ version is shown in part (a), with the numbers that are assigned to the words for statistical analysis. The numerical ‘numbers only’ scale shown in part (b) is sometimes presented to consumers and is labeled at the ends and sometimes at the midpoint.

It is not surprising that Peryam and Giradot⁸ and later Peryam and Pilgrim⁹ when introducing the scales in the food science literature discussed them in terms of their comparative use; they were scales of liking from which preference was to be inferred. Therefore, the words on the scale were to be considered as points on a continuum rather than categorical discrete data. Using the responses for the ‘words only’ scale in an absolute fashion, as with menu planning, was considered a special case.⁹

Yet, the point is not trivial. If the goal of the measurement is preference, then the experimental protocol should make comparisons between the products as simple as possible. What is often ignored during the design of scaling protocols is the effects of rapid forgetting. Therefore, consumers should be allowed to re-taste stimuli to check their hedonic assessments as much as required and to alter their scores accordingly.

Notwithstanding, it is sometimes argued that in the interests of realism, products should be tested serially monadically (each product assessed once with no re-tasting or reference to previous scores). The idea is that in the purchasing situation consumers do not taste a set of products in the shop and then choose the one they prefer. They simply choose one product and buy it. This is quite true but it can also be argued that consumers choose products, albeit one at a time, based on their memory of comparisons with previous tastings of the available products.

The idea of the serial monadic protocol is to promote absolute judgements that are not influenced by comparisons with other products. The idea is to eliminate context effects. Consequently, using a serial monadic protocol for comparative measures of preference would be to measure preference as interfered with by the effects of forgetting. It will introduce errors whereby a more preferred or more intense stimulus can be given a lower score than a less liked or less intense stimulus, as discussed in the next section.

HOW DO CONSUMERS ASSESS SCALES: WHAT ARE THE COGNITIVE STRATEGIES?

As mentioned above, the 9-point hedonic scale is a liking scale used to measure preference. The nine verbal categories on the ‘words only’ scale are generally assigned numbers from 1–9 and the responses to the verbal categories are treated as responses

to numerical values along a preference continuum, namely a ‘numbers only’ scale. The developers of the scale admit that the numbers produced on this ‘numbers only’ scale are not equally spaced^{8,9} and more liked ranked data, but treating them as equally spaced is justified by the need and does not cause any major trouble. In the light of this, it is useful to consider cognitive strategies or decision rules, used by the brain, for some background on the use of category scales. There are two rival models for cognitive strategies: Zwislocki’s absolute model^{15,16} and Mellers’s relative model.^{17,18} They will be discussed here using ‘numbers only’ scales as examples; examples of ‘words only’ scales are discussed below.

Zwislocki’s absolute model

The absolute model hypothesizes that a stimulus to be assessed for intensity or liking, is compared to a set of internal exemplars stored in the brain, each representing a given degree of intensity or liking and to which is assigned a numerical value (see Fig. 2a). For example, product ‘A’ is perceived as being more intense or liked more than exemplar 2, but less intense or liked less than exemplar 4. Accordingly, it is given a score of ‘3’. A similar argument applies to product ‘B’ (4) and product ‘C’ (8). This model demands a protocol whereby the judgement made for one stimulus is not affected by judgements made for other stimuli. There should be no context effects. When consumers are assessing product ‘B’, they should only be attending to the exemplars and their numerical values. They should not be remembering product ‘A’ because comparison with its sensation and assigned score might bias the assessment of product ‘B’. Basically, product ‘A’ should be forgotten. In the same way, the memory of products ‘A’ and ‘B’ should not affect assessment of product ‘C’. To attain these conditions, to avoid any context effects, a monadic protocol would be suitable. Each product would then be assessed on a different day or in a different week. Yet, this is usually not practical so the compromise serial-monadic protocol is used, where products are assessed one after another but once the assessment has been made the product and access to the responses given to other stimuli are removed. This presentation protocol is a standard procedure for the 9-point hedonic scale, although it is more suitable for calibrated laboratory instruments, where calibration

establishes the numerical exemplars and there are no context effects.

Mellers's relative model

The alternative relative model (see Fig. 2b, top row of numbers) assumes that the score given to a product depends on the scores given to the other products. For example, product 'X' is only slightly less intense or less liked than product 'W'. Therefore, the score assigned to it will be only slightly less (8 vs. 9). However, product 'X' is certainly more intense or more liked than product 'Y', which accordingly is given an appropriately lower score (5). Product 'Z' is much less intense or less liked and is assigned an appropriately even lower score (1). Another way of describing this is that the process is akin to ranking and using the scale numbers to describe the spacing between the ranks. Because all consumers do not use rating scales in an identical manner, other consumers may have given the scores 7, 6, 4 and 1 or 8, 7, 5 and 2 (see Fig. 2b, bold numbers, 2nd and 3rd rows). They may have used different numbers, but they are all trying to describe the same picture: 'W' came first with 'X' a close second; 'Y' came third but a little further behind, while 'Z' came fourth and far behind. Furthermore, unlike the absolute model, there is no implication that a score of '9' for product 'W' represents a greater intensity or a greater degree of liking than score of '7' or '8' given by other consumers.

Which models apply to 'numbers only' and 'words only' scales

The question becomes: which model is correct for consumers using hedonic scales? For purely numerical scales, context effects^{19–23} provide evidence for the relative model. Indeed, Mellers^{17,18} used such effects to argue against Zwislocki.^{15,16} Lawless and Malone²⁴ also used context effects to argue that intensity scaling was relative. Poulton's²⁵ stimulus equalizing bias has also been used as an argument for a relative model.^{24,26,27} This bias refers to the fact that stimuli that are relatively similar and stimuli that are relatively different tend to be spaced in the same way across the whole length of the scale.

Still further evidence comes from studies on forgetting. The scores given to products during rating and the exact sensation elicited by those products can be forgotten within seconds. Rank-rating is a protocol²⁸ whereby consumers are able to re-taste stimuli as often as desired and review and alter their scores as often as is required; this allows them to avoid giving inappropriate ratings caused by forgetting. One way of achieving this is simply to print the numbers (1–9) on a suitably large visible scale and require consumers to place products on or in front of the appropriate numbers. With the ability to re-taste the products as much as required, they can be freely moved up and down the scale, until the final 'picture' of their relative intensities or degrees of liking is represented. For example, they can check whether they really did give a higher score to a stimulus they liked more. When this protocol is compared with a serial monadic protocol, where re-tasting and the monitoring of scores are not allowed, the rank-rating protocol unsurprisingly elicits fewer errors due to memory loss, which supports the fact that a relative model best fits the cognitive strategy for a 'numbers only' scale.^{27,29–32} These results and their relevance to the absolute and relative models have been reviewed.³¹ Therefore, in view of these results, it would be wise to recognise that the 'numbers only' scale is best described by a relative model and from this, it follows that a rank-rating protocol which allows re-tasting and monitoring of the scores given, is appropriate where possible.

On the other hand, for the 'words only' version, there is evidence that the cognitive strategy is closer to the absolute model.^{26,27}

Here, the words act as linguistic exemplars, which may stay constant for a given consumer for the length of time that measurements are being made. The task of the consumer is merely to categorise each product as liked or disliked to different degrees. It is similar to assessing products for menu planning. Regarding the categories or exemplars used on the scale, it would not be expected that these word-generated exemplars would stay constant. Also, they would not necessarily express the same degree of liking between consumers or for the same consumer over a longer time.

WHAT SORT OF NUMBERS ARE GENERATED BY 'NUMBERS ONLY' SCALES?

Types of numbers

Regarding the types of numbers generated during scaling, Stevens^{33,34} devised a language to allow adequate description of the status of numbers. They can be categorized as nominal (numbers used as names), ordinal (ranks), interval (numbers equally spaced without a true zero) or ratio (equal spacing and a true zero). With this scheme in mind, it is possible to consider evidence for a consumer's behavior during scaling.

Numbers obtained from category scales

For a 'numbers only' scale, it is common for a consumer to be reluctant to use the end of the scale. In psychology, this is called an end effect. It is as though it appears cognitively more difficult to pass from 8 to 9 on a 9-point scale, than it is to pass from 4 to 5. Somehow the journey requires more cognitive effort; it is as though the distance feels greater. It is as if the spacing between 8 and 9 is bigger than that between 4 and 5. It is as if although numbers *per se* are equally spaced, psychologically numbers generated by a rating scale are not; thus, we do not have an interval scale. If one consumer spaces numbers in one way, it is unlikely that a second consumer would space numbers in exactly the same way. This is not an insurmountable problem; it is easily circumvented by using an appropriate experimental design.

A second line of evidence comes from the process of fitting d' values to rating data.³⁵ Estimates of distribution means and decision boundaries are obtained by the method of maximum likelihood, using an extension of the method discussed by Dorfman and Alf³⁶ and Ogilvie and Creelman.³⁷ From this method, the variance-covariance matrix of the parameter estimates can be obtained so that statistical tests on the estimates (e.g. significance of differences between d' values) can be conducted. Software has been written to facilitate this (IFPrograms, Institute for Perception, Richmond, VA, USA). As expected, the boundaries between adjacent categories do not turn out to be equally spaced, because some scores are used more frequently than others. These boundaries are equivalent to a set of β -criteria^{38–40} in signal detection theory^{41,42} and as such are variable between consumers and over time. This phenomenon is called boundary variance.⁴³

Considering these facts, it would seem that data generated by the 'numbers only' version of the 9-point hedonic scale are at least ordinal. Yet, they provide more information because the consumers use the numbers to represent the spacing between the ranks. However, this does not make them as good as interval data. Consequently, the data provided by the 'numbers only' version of the 9-point hedonic scale are better than ordinal but not as good as interval.

Regarding the 'words only' version of the 9-point hedonic scale, Peryam *et al.*¹³ reported that because of its construction, the

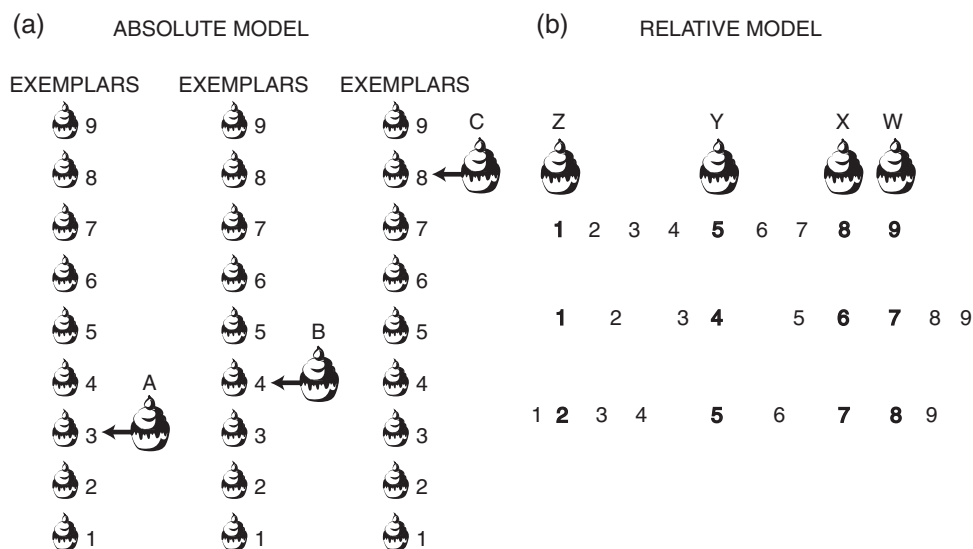


Figure 2. Models for cognitive strategies used in scaling. Part (a) represents the absolute model whereby scores of intensity or degree of liking are generated by matching them to exemplars stored in the brain. Part (b) represents the relative model whereby scores for a product are generated relative to the scores given to other products, to provide an overall picture.

data generated are only ordinal. They acknowledged that fitting numbers to such 'words only' data was not statistically correct. Other researchers concerned with the development of the scale reached the same conclusion (e.g.⁴⁴).

ON ASSIGNING NUMBERS TO THE 'WORDS ONLY' 9-POINT HEDONIC SCALE

Development of the 9-point hedonic scale

At this point, it is worth considering how the U.S. Army Quartermaster Food and Container Institute developed the 9-point hedonic scale as an aid for menu planning in army canteens, as far back as 1949. This has been described by several authors.^{2,10,12,13} After introduction of the scale,⁸ developmental work was performed by Jones and Thurstone⁴⁴ and Jones *et al.*⁴⁵ using scaling techniques developed by Edwards.⁴⁶ For this, 834 or 829 soldiers, depending on which report you read,^{44,45} were given 51 words and phrases that could be used as hedonic descriptors. The hedonic strengths of these words and phrases were rated on a numerical bipolar category scale, ranging from '-4' (greatest dislike) through '0' (neither like nor dislike) to '+4' (greatest like). The assumption was made that the grand total of scores from all words and phrases were normally distributed along the scale. The center of this distribution did not fall exactly on the zero value of the scale. The raw scores (-4 to +4) were then converted to z-scores. Normal distributions for these z-scores, associated with each individual word or phrase, were noted and their means and standard deviations were computed. Distributions with small standard deviations, having little overlap, indicated a lack of ambiguity for the hedonic strengths of these words and phrases. Those that were suitably unambiguous were selected for the hedonic scale. This resulted in an 11-point scale, which in the days of typewriters, did not fit on the paper, so the scale was reduced to a 9-point scale. There was no pretence that the words and phrases chosen were equally spaced along the scale, resulting in ordinal data. This was later supported by experimental work by Stroh.⁴⁷

Peryam and Girardot⁸ suggested assigning the numbers 1 to 9 to the nine verbal categories in the scale and the data be treated quantitatively, calculating means, standard deviations and significance of difference between means using analysis of variance. Interestingly, Peryam and Pilgrim⁹ later stated that the data were no better than ranks. Yet, both sets of authors stated that although the validity of using such statistical methods might be questioned by some statisticians, such analysis was required and could be justified on practical grounds, until appropriate techniques became available. Since then, assigning numbers to the words and phrases in the traditional 'words only' version of the 9-point hedonic scale, with subsequent parametric statistical analysis, has become a generally unchallenged routine; the numbers are commonly treated as at least interval data, derived from a normally distributed population.

A cautionary note

Yet, as statistical analysis and modeling becomes more complex, it is as well to remember the assumptions involved. Merely writing numbers next to categories does not make the categories numerical. It is merely using alternative names for the categories. For example, if the category 'like extremely' is assigned the number '9', the category has not become a number; it has merely been given a new name: 'nine'. There is a further point. As discussed above, data from a 'numbers only' scale are better than ordinal but not as good as interval. The numbers assigned to a 'words only' scale have the same property as the words; they are merely ordinal. Giving it a category new name does not magically turn it into a number, which comes from a normal distribution and is subject to parametric statistical analysis. Any such analysis is liable to be approximate, because the ordinal nature of the assigned numbers would distort any normal distribution. This position was disputed by Stone and Sidel,¹⁴ who claimed that the 9-point hedonic scale does not violate the normality assumption. In support of their position, they showed cumulative data from the scale that was represented by a sigmoid-shape curve. Yet, distorted as well as regular normal distributions can yield sigmoidal functions.

The simplest illustration that the assigned numbers are merely new labels is that if 'eight' is subtracted from 'nine', the answer is 'one'. Yet, if 'like very much' is subtracted from 'like extremely', the answer is hardly likely to be 'dislike extremely'.

Therefore, in summary, it must be said that there is no problem with the 9-point hedonic scale *per se*. It has the advantage of being convenient and easy to use, while many companies have records of its use, for comparison with past products going back for many years. The problem arises with those who are not aware of the true nature of their scaling data and the approximate nature of their parametric analyses. For parametric statistical analysis, interval data are required. This means that the statistical analysis of numbers assigned to the words on the scale will be approximate. This will not necessarily cause a great deal of harm, as long as the users are aware of this and can interpret the data accordingly. There is no problem in working with approximate data as long as the user realises that it is approximate. The potential for problems occurs when sensory professionals treat their data as though they were produced by a calibrated laboratory instrument.

With this in mind, it is worth contemplating the words of Oppenheim⁴⁸ from his textbook on consumer testing: 'The use of ratings invites the gravest dangers and possible errors, and in untutored hands the procedure is useless. Worse, it has a spurious air of accuracy, which mislead the uninitiated into regarding the results as hard data.'

'WORDS ONLY' AND 'NUMBERS ONLY' 9-POINT HEDONIC SCALES

Versions of the 9-point hedonic scale

Generally, the 'words only' version of the 9-point hedonic scale is presented to consumers and the numbers 1–9 are attributed to the words for subsequent statistical analysis. However, recently, some researchers have been presenting 'numbers only' 9-point scales to consumers with the ends of the scales representing liking most and liking least (or disliking). These labels can be communicated either verbally or by labeling the ends of the scale. Unfortunately, authors sometimes fail to report whether they present the 'words only' or 'numbers only' version of the scale.

Inspecting the literature, it can be seen that there are plenty of examples in the way these variations are presented. For example, some authors using the 'numbers only' scales are precise in the way they describe their scales,^{49–51} while for others, it can be inferred.⁵² The same is true for the 'words only' scale, where some authors are precise in their descriptions or have given personal communications (for example:^{53–57}), while others have referred to textbooks (for example:⁵⁸). In other reports, there is room for ambiguity (for example:^{59–62}).

'Words only' scales and 'numbers only' scales are not equivalent

Using the 'words only' version of the 9-point hedonic scale, food products that were liked, would be expected only to be rated in the top half of the scale (like slightly to like extremely). For a 'numbers only' scale, it would be expected that products would be rated across the whole length of the scale, with no implication of dislike for scores ranging 1–4. As mentioned above, this tendency is described as Poulton's²⁵ stimulus equalizing bias. Accordingly, it would be possible for some foods all to be rated as 'like very much' (all with derived scores of '8') on the 'words only' scale, while they might stretch along the whole length of a 'numbers only' scale. The

numerical mean values derived from the two scales would not be the same.

This has been demonstrated experimentally using both rank-rating and serial monadic protocols.^{26,27} Interestingly, for the serial monadic protocol, some consumers made errors in that they gave relatively lower hedonic scores to foods that they liked relatively more. This was despite the fact that consumers had earlier ranked the products in order of liking and a duplicate set of this ranking was visible, while the consumers were making their ratings; they agreed that the errors were caused by forgetting the ratings they had given and being unable to make checks. Such errors would reduce the power of this scaling protocol.

Cognitive strategies were also studied using Poulton's²⁵ stimulus equalizing bias. This indicated that consumers using the 'numbers only' scale used a relative cognitive strategy, while with the 'words only' scale, consumers used a cognitive strategy that was mostly but not totally absolute in nature.

In summary, researchers reporting the use of a '9-point hedonic scale' should specify whether consumers were presented with the traditional 'words only' scale or the alternative 'numbers only' scale. These two protocols do not give equivalent data so that direct comparison is not possible. Accordingly, any other presentation forms of the 9-point hedonic scale like a structured scale² or a 'box scale'^{63–65} need to be carefully described and an illustration might sometimes be required.

VARIOUS ANALYSES FOR THE 9-POINT HEDONIC SCALE

A product might be developed to fit a concept, which in turn was developed by a marketing department. Alternatively, a product might be developed to challenge a market leader or to decrease (e.g. salt, fat) or add particular ingredients (e.g. omega 3 fatty acids, fibre). For such product development, it is important to produce a product that consumers will like; after that, it is up to the marketing department to turn this into a product that consumers will buy. After the usual preliminary testing for product development, it becomes time to perform some form of hedonic testing on a sample of appropriate consumers.

'Numbers only' analysis

With a 'numbers only' scale, the new product could ideally be tested against the market leader in the same product niche, as well as, perhaps, a market middle seller and the market loser. If the new product gets a score close to the market leader, it is good news; the new product is liked as much or more than the market leader and it does not matter whether the market leader's score was '9', '8', or '7' as long as the new product's score is close. A 'numbers only' scale is relative not absolute like the 'words only' scale. On the other hand, if the new product has a score closer to the market loser or even the market 'middle', the news would not be good. The product development might be cancelled.

It can be argued that the clever thing about this approach is that the market research for the new product has already been done. The product is compared with foods whose market performance is known. Because these are measures of liking, the tasting would be performed 'blind'.

Should the new product be liked more than or as much as the market leader or even closely behind the market leader, the decision may well be to give the product to the marketing department, for them to develop an appropriate marketing strategy.

At this point, the sensory department may no longer be involved. However, in sensible companies, where marketing and sensory sciences are part of the same department or at least work closely together, a similar approach could be advised to test the efficacy of the marketing strategy. The same protocol could be used, except that a purchase intent scale would be used instead of the hedonic scale. In this case, however, the products would not be tasted 'blind'. They would all be presented in their cartons, along with any marketing messages and advertisements that were used in their marketing campaigns. It is quite possible in this case that a well-liked new product would not score as highly as the market leader. It might even be closer to the market loser. If this were so, the marketing department would have been warned that their planned marketing strategy was likely not to be effective. Accordingly, because of the use of suitable sensory testing, the company would have been saved from wasting money on an inadequate marketing strategy. A better marketing plan is required.

'Words only' analysis

Regarding the 'words only' hedonic scale, it is not likely that a company whose usual protocol is to attribute 'numbers' to the words on the scale, would want to change their analysis. A company may have records that show that mean scores of these attributed 'numbers' have some predictive ability. A sensory professional may have gained sufficient experience to make meaningful predictions from these mean scores. If this were truly the case; stay with the method. If it is not broken, don't fix it.

However, the experienced sensory professional might eventually retire at some time or a company may not have extensive records. Therefore, a method that represents the data more clearly would be suitable. A standard method for analyzing data from verbal categories is to use a simple histogram. Yet, this was suggested before for the 9-point hedonic scale. Peryam and Girardot⁸ in 1952 suggested that using the percentages of responses that fell into each of the verbal categories, was statistically more respectable, and as useful as attributing numbers to the categories and computing a mean. However, this approach was not adopted. Later, Peryam *et al.*,¹³ in their discussion of the appropriate analysis for the 'words only' scale, considered this approach but rejected it as "much more cumbersome to report and discuss than if a single index (the mean) were used." They went on to point out difficulties with analyzing such data and the fact that a variance could not be accurately determined. Yet, with the data they were using, their computation of variance would hardly be accurate. Furthermore, with computers, histograms are not cumbersome and give more information than a mean.

For example, consider the two histograms shown in Fig. 3. Both have the same mean (5.8) which would be interpreted as 'like slightly'. Yet, the histograms tell very different stories. The top histogram indicates that the product A is fairly innocuous. The highest frequency falls in the 'neither like nor dislike' category; it offends nobody. Yet, enough consumers like it 'slightly' or 'moderately' to give it a mean score in the 'like slightly' category. This is typical of products which are designed to appeal to a wide range of consumers and which might be considered safely bland. The marketing decision may be to put the product on the market to achieve high market share and the marketing strategy would reflect this. On the other hand, the bottom histogram for product B has the same mean value but its distribution is quite different; it indicates serious segmentation. It is not bland and inoffensively safe like product 'A'. Some consumers appear to like it while others dislike it. This would be more suitable as a niche product and

should be marketed differently from product 'A', if at all. It elicits many 'like very much' and 'like extremely' responses, none of which are elicited by product A. Statistical analysis is not really needed for these products; the histograms tell the whole story. However, if a statistical analysis is required, suitable use of binomial type comparisons can be arranged with a little creativity.

If a company or its experienced sensory professional insist on calculating a 'mean' and 'standard deviation' from the numbers assigned to the words on the scale, it is as well to add histograms to the analysis, to avoid any misinterpretations. Also, the histogram picture is easy to understand for those in the business community.

HEDONIC RANKING AND A SUGGESTED ANALYSIS

Advantages of hedonic ranking

A different approach to hedonic measurement is hedonic ranking. There are some advantages to hedonic ranking. Although rating scales are not problematic for a large proportion of the population, they can cause problems for some elderly consumers, as well as those with reading difficulties or simply less at ease with mathematical tasks. Ranking, however, is an activity which most consumers find easy because it is something at which they are practiced. From childhood, people have ranked things. Little boys may rank their favorite football team, second favorite, third favorite, etc. Their sisters may rank their favorite pop star, second favorite, third favorite, etc. Accordingly, it is sensible to test consumers using a method with which they are familiar. A second advantage to ranking is that it encourages judges to re-taste stimuli that they may have forgotten; such double-checking avoids errors due to forgetting. A third advantage is that it uses human behavior as its data source rather than the imperfect numerical estimates obtained from rating.

Analysis of ranked hedonic data: John Brown's R-Index

Unlike ranking for intensity, where protocols tend to be forced choice, tied ranks are generally allowed to enable expression of 'just about equal' liking. One obvious numerical measure of hedonic preference for ranking is to take mean ranks. Yet, these are context dependent. Another measure is to use John Brown's R-Index.⁶⁶ This is a probability measure that was first used in food science for difference testing. It has been reviewed in detail⁶⁷ and also more briefly.⁶⁸ Tables of significance are also available.⁶⁹ The R-Index has been applied in various difference testing studies (e.g.⁷⁰⁻⁷²).

The R-Index has also been used as a measure of degree of preference.⁷³⁻⁷⁵ Yet, it is the use of the R-Index made by Pipatsatayanuwong *et al.*⁷⁶ as a substitute for hedonic scaling that is of interest here. They used a ranking technique to assess consumers' liking for different temperatures of coffee. They used 'hedonic R-Index' measures to represent the degree of difference in liking between the rank order of preferred coffee temperatures.

The use of hedonic R-Indices gives the same information as would be obtained from mean scores on a hedonic scale; this is illustrated in Fig. 4. Consider the case, shown in the figure, where four products, 'R', 'L', 'M' and 'N' were given mean hedonic scores of '8', '5', '4' and '2', respectively. This is interpreted as product 'R' being liked considerably more than product 'L'. Products 'L' and 'M' are liked more similarly and product 'M' are liked rather more than product 'N'. This information gives us the rank order of the products, with the mean scores describing the spacing between the ranks.

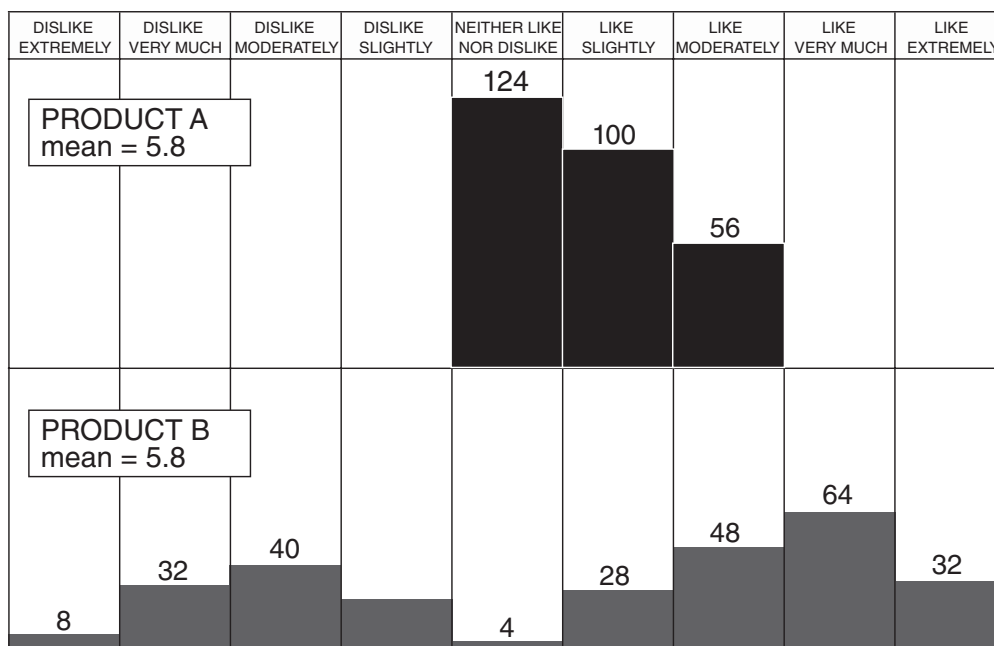


Figure 3. Two histograms representing data from the traditional 9-point hedonic scale. Even though products 'A' and 'B' have the same mean values, their distributions are very different. Product 'A' is fairly inoffensive while product 'B' is either liked or disliked.

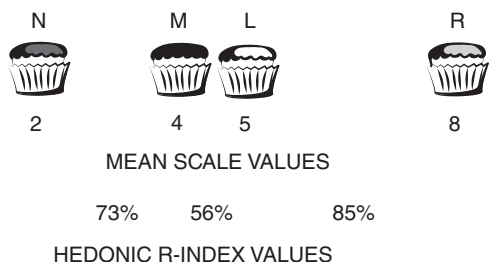


Figure 4. Mean hedonic scale values calculated for products 'N', 'M', 'L', 'R'. These give the rank order of liking for these products, with the mean scores representing the spacing between the ranks. The hedonic R-Index values give the same information, using probabilities of consumers choosing the more preferred product over the less preferred product.

Hedonic R-Index values give the same information, yet, in a different way. R-Indices in Fig. 4 indicate that 85% of the consumers tested preferred 'R' to 'L'. Only 56% preferred 'L' to 'M', while 73% preferred 'M' to 'N'. A hedonic R-Index of 50% indicates equal overall preference for two products, while 100% indicates that all consumers have the same preference.

COMPUTATION OF THE HEDONIC R-INDEX

The method of computation

Although the R-Index computation has been described before for difference tests,^{67,68} it has not been described in detail for the hedonic R-Index. Accordingly, consider Fig. 5. Four fruit flavored products: apricot (A), banana (B), cherry (C) and nectarine (N) are ranked for liking by 10 consumers. Obviously, this is far too few, but it will be enough to illustrate the computation. Consumer #1 can be seen to like 'A' the most and 'N' the least. Consumer #2 likes 'B' the most, while consumer #5 likes 'C' the most. All 10 consumers like 'N' the least or dislike 'N'. The frequencies of coming 1st, 2nd, 3rd or 4th are given for each product in a response matrix (see Fig. 5). It is from this matrix that hedonic R-Indices can be computed.

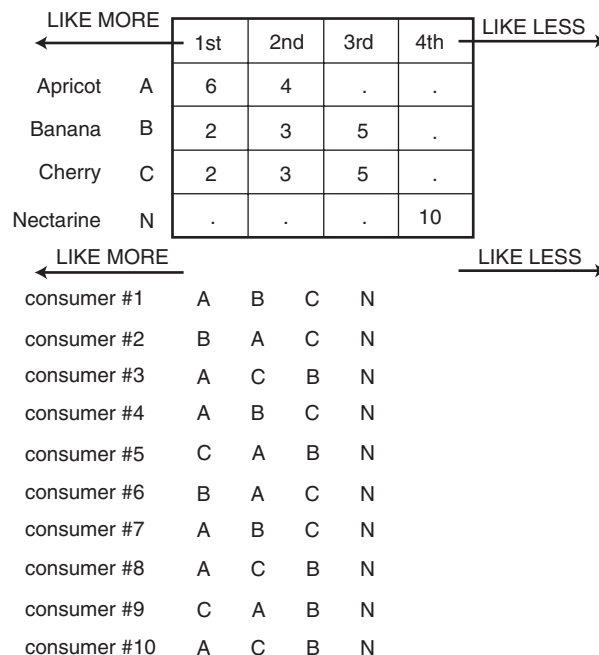


Figure 5. Rank order of liking for fruit flavored products: apricots (A), banana (B), cherry (C) and nectarine (N), using 10 consumers. A summary of their rankings is given in an R-Index matrix.

From the matrix, consider the computation for the preference of product 'A' over product 'B'. Imaginary comparisons are made between all ten 'A' products and all ten 'B' products. This gives a total of 100 (10 × 10) comparisons. In such comparisons, a product ranked as more liked would be judged to be preferred to a product ranked as less liked. The number of comparisons from 100 in which 'A' is preferred to 'B' is the R-Index. It can be seen as the percentage probability of these consumers preferring 'A' to 'B' or alternatively the percentage of consumers in the sample who preferred 'A' to 'B'.

	← LIKE MORE	1st	2nd	3rd	4th	LIKE LESS →
Apricot A		6	4	.	.	
Banana B		2	3	5	.	

Prefer A to B	30	+	18	=	48	
Prefer A to B	20				total	68
Matrix tie	12					
Matrix tie	12				total	Matrix tie 24
Prefer B to A	8					

Grand total	
68	+ 24
=	100

Figure 6. Summary of the hedonic R-Index computation, representing preference for the apricot flavored product (A) over the banana flavored product (B).

Using the hedonic R-Index matrix shown in Fig. 6, consider the six 'A' products ranked 1st being compared to the five 'B' products ranked 3rd and the three 'B' products ranked 2nd. Because the 'A' products were given ranks indicating greater 'liking' than the 'B' products, it can be said that the 'A' products would be preferred. This gives $(6 \times 5 = 30) + (6 \times 3 = 18) = 48$ estimated preferences for 'A' over 'B'. Now consider the four 'A' products ranked 2nd and the five 'B' products ranked 3rd. Using the same logic, these $4 \times 5 = 20$ comparisons would indicate, once again, preference for 'A' over 'B'. Thus, so far, we have estimated a total of $48 + 20 = 68$ preferences of 'A' over 'B'.

Now consider the six 'A' products ranked 1st and the two 'B' products also ranked 1st. In this case, we cannot decide which of these $6 \times 2 = 12$ comparisons indicate a preference for 'A' or 'B'. Both 'A' and 'B' were given 1st place, but that does not necessarily mean that they were liked equally. Thus, these 12 comparisons will be scored provisionally as a 'matrix tie'. Likewise, the four 'A' products that came 2nd and the three 'B' products that also came 2nd produce $4 \times 3 = 12$ more matrix ties. Finally, the two 'B' products that were ranked 1st when compared with the four 'A' product ranked 2nd would produce $4 \times 2 = 8$ preferences for 'B' over 'A'.

Overall, there are 68 preferences for 'A' over 'B' and 8 for 'B' over 'A'. The next task is how to treat the $12 + 12 = 24$ matrix ties. A tie in the matrix could occur either because 'A' is slightly preferred to 'B' or *vice versa*. As there is no way of estimating the relative numbers of these preferences, the easiest procedure is to split them equally. This gives $12 + 68 = 80$ preferences for 'A' over 'B' and 20 preferences for 'B' over 'A'. Another way of expressing this is that the probability of these consumers preferring 'A' to 'B' is 80% or that 80% of the consumers prefer 'A' to 'B', while preference for 'B' over 'A' is 20%.

Next, we can consider preference for 'A' over 'C'. The numbers for this computation are exactly the same as for calculating the preference for 'A' over 'B'. Accordingly, the preference for 'A' over 'C' is 80% and for 'C' over 'A' is 20%. Considering the preference for 'B' over 'C', it can be seen from the matrix that the rankings are identical. Therefore, a calculated hedonic R-Index will equal 50%. Considering comparisons with product 'N', all consumers preferred 'A', 'B' and 'C' over 'N', resulting in an R-Index of 100% (Fig. 7).

Hedonic ranking allows ties. Should two products tie and be ranked in 2nd place, the usual statistical rule is employed and they are regarded as tying for 2nd and 3rd places. They both get the rank of '2.5'. Thus, an extra column is introduced into the matrix

		← LIKE MORE	1st	2nd	3rd	4th	LIKE LESS →
Apricot	A		6	4	.	.	
Banana	B		2	3	5	.	
Cherry	C		2	3	5	.	
Nectarine	N		.	.	.	10	

HEDONIC R-INDEX

$$\text{Prefer A to B} = 80\%$$

$$\text{Prefer A to C} = 80\%$$

$$\text{Prefer B to C} = 50\%$$

$$\text{Prefer A to N} = 100\%$$

$$\text{Prefer B to N} = 100\%$$

$$\text{Prefer C to N} = 100\%$$

Figure 7. Preferences between all products as calculated from the hedonic R-Index matrix.

between the 2nd and 3rd columns representing the rank of '2.5'. The hedonic R-Index computation is carried out in the same way with the addition of an extra column. In reality, with large samples of consumers, it would be anticipated that several of these 'tie' columns would be necessary. However, this does not affect the computational method.

A more realistic example

Consider a more realistic hedonic R-Index matrix for 400 consumers, assessing a prototype product against the market leader, the market number 2 and a market loser (see Fig. 8). Because there are so many consumers, the number of consumers who report that they liked some of the products 'just about the same' and accordingly produce tied ranks, is likely to be substantial. Therefore, this matrix has extra columns for products that tied for 1st place, 2nd place and 3rd place and were accordingly assigned the ranks of 1.5, 2.5 and 3.5, respectively. From the matrix, the calculated hedonic R-Indices indicate that only 52% of the consumers preferred the market leader to the prototype, which itself was preferred to the market number 2 by 82% of the consumers. The market number 2 was preferred to the market loser by 67% of the consumers. Regarding significance, Bi and O'Mahony's⁶⁹ tables indicate that for 400 consumers, a hedonic R-index of 53.99% or greater is significant at $P < 0.05$. This means that there was not a significant preference for the market leader over the prototype, which was greatly preferred to the market number 2. In this case, the company would be very likely to consider seriously continuing with the programme of launching the prototype on the market.

R_{MAT} and R_{JB}

The computation of R-Indices using the matrix is only one way of calculating this index. John Brown^{66,67} had an alternative method. He simply counted the number of times that a given product preceded a second product in the rankings. For example, it can be seen from Fig. 5 that 'A' was ranked in front of 'B' 8 out of 10 times giving an R-Index of 80%. 'B' was ranked ahead of 'C'

RANKS	1	1.5	2	2.5	3	3.5	4
MARKET LEADER	280	79	41	0	0	0	0
PROTOTYPE	277	50	40	33	0	0	0
MARKET NUMBER 2	82	0	119	40	79	77	3
MARKET LOSER	0	0	152	8	80	75	85

Figure 8. A more realistic hedonic R-Index matrix using 400 consumers, to compare a prototype product to products already on the market.

50% of the time, while all were ranked ahead of 'N' 100% of the time. These are the same values as were computed from the matrix (see Fig. 7). R-Index values computed from the matrix are designated ' R_{MAT} ', while those computed by John Brown's method are designated ' R_{JB} '. By happy coincidence, in this example, both sets of values were identical, but this is not always the case, so it is as well to check. Neither value is more correct than the other; they are simply different. ' R_{MAT} ' has the advantage of including the degree of difference between two products, because they are placed in different columns in the matrix, which in turn affects the overall R-Index value. On the other hand, R_{JB} only considers how many times one product precedes another in the ranking, regardless of the rankings of any other products; it is thus more context free.

It is unlikely that sensory professionals married to the 9-point hedonic scale will abandon it in favor of R-Index hedonic ranking. This is despite the fact that R-Indices are derived from human behavior rather than the imperfect numbers generated from rating scales and that R-Indices, being probability values, are susceptible to parametric statistical analysis. However, R-Indices can easily be computed from the data obtained by hedonic scaling by simple conversion into ranks. This is especially easy using R_{JB} . Giving a percentage of consumers who prefer 'A' to 'B' is useful information that helps clarify the meaning of the relationship between their two mean numerical scores on the 9-point hedonic scale. It is certainly a useful tool for communication with those outside the sensory community. When asked, what it means when one product is given a mean score of '8' and another is given a mean score of '5', it can be explained that 85% of consumers preferred the former product (see Fig. 4). It can also be used to bring new meaning to past records. Thus, although the present authors see the advantages of R-Index ranking, they do not anticipate that sensory professionals will rapidly switch away from the 9-point hedonic scale after sixty years of tradition. At present, it can be seen as a useful clarifying supplement to data obtained by the 9-point hedonic scale and which may eventually precede it.

However, for disciplines other than food science, which might be free of the compulsion to use the 9-point hedonic scale, ranking with an R-Index is a suitable method to adopt as the primary method for measures of liking, because of its simplicity for those being tested. It is suitable for any discipline that makes behavioral measures of liking.

The R-Index ranking technique has been applied to more measurements than hedonic scaling. It can be applied to the measurement of abstract consumer concepts. It has been applied, using simple ranking, for measuring how strongly visual appearance elicits the feeling of refreshingness for toothpaste.^{77,78}

SOME PROBLEMS WITH THE 9-POINT HEDONIC SCALE

Preferences among equally liked products

Simone and Pangborn⁷⁹ tested over 2000 attendees at the California State Fair, to determine preference for canned cling peaches, with and without added citric acid. Consumers responded using the 'words only' 9-point hedonic scale. They reported that 'participants were frustrated when they liked the two samples equally, but had a preference for one, and could not indicate this on the response sheet.' However, they did not report the proportion of participants who had made this complaint. Villegas-Ruiz *et al.*⁸⁰ investigated this, using three strawberry yogurts and the same 'words only' scale. Of their 116 consumers, they noted that 32 (28%) gave identical responses for two of the yogurts, while one consumer gave identical responses for all three. In all cases, all consumers reported that tied scores did not represent equal liking.

There are several ways round this problem. The categories on the scale could be split into subsections to provide ways of representing preferences. Another alternative would be to require consumers simply to rank the foods, allowing ties and use a hedonic R-Index analysis.

Cross-cultural use of the 9-point hedonic scale

Yeh *et al.*⁸¹ compared the use of the 'words only' scale with American, Chinese, Korean and Thai consumers. They found that Chinese, Korean and Thai consumers used a smaller range of the scale than Americans. They hypothesized that these smaller ranges were due to an Asian enculturation to avoiding extreme responses. They also suggested an Asian tendency to be polite by not expressing negative responses. Yao *et al.*⁸² compared responses among Americans, Japanese and Koreans on a 'words only', 'numbers only' and a third scale using both words and numbers. For all three scales, they found that Japanese and Koreans used a smaller range of scores than the Americans. Subjective responses indicated that the smaller range for the Japanese was avoidance of the lack of politeness in expressing extremes. However, for the Koreans, it was the avoidance of the Korean translation for the 'like extremely' and 'dislike extremely' categories, because the translation of 'extreme' into Korean had a meaning with excessively strong semantic weighting.

With globalization, it becomes important for multinational companies to be able to compare consumer acceptance for the same product in different countries. For this, hedonic measures will be one of the tools required. Accordingly, it would be necessary to use the same testing protocol and the same scale in different countries. Yet, these results indicate that the Asian groups tested were using a shorter scale than the Americans, so that direct comparison becomes problematic. A universal scale, used in the same manner in different countries is required. One solution is to use hedonic ranking. The length of the 'scale' will be the same in all countries, namely the number of stimuli being tested. Feng⁸³ used hedonic ranking to compare responses between mainland Chinese and American consumers. She also used 'words only' and 'numbers only' hedonic scales. As with the former research, she found that Americans used a wider range scores for both 'words only' and 'numbers only' scales. However, this effect was obviously absent for hedonic ranking, demonstrating its utility.

Too many products to rank

Finally, if there are too many products to assess, as can happen with a product optimization exercise, products can be divided into

separate groups, for example: high, medium and low. Each group can then be ranked separately and then the three groups joined to produce one long ranking. This cannot be done for scaling data. In psychology, this procedure is called 'two-stage' ranking. Another possibility would be to use the 'words only' hedonic scale to divide the products into separate groups. Again, each group can be ranked separately and later joined for a complete set of ranks. Birch *et al.*⁸⁴ used a similar procedure, creating groups from a 3-point 'smiley-face' scale.

CONCLUDING SUMMARY

It can be sensibly argued that if a hedonic scale is only being used to get an idea of the relative liking for various products, so that the most liked can be selected for further development, it probably does not matter which hedonic scale is used. Ratings derived from all hedonic scales should give the same rank order, other things being equal. So it may be asked why the present authors are being so pedantic. This is because sensory science is at a point where data from hedonic scales are being entered into complex multivariate programs used for statistical analysis and modeling. Because of this, it is time to pause and consider whether sometimes the questionable nature of the numerical data from rating scales that we are using in these programs, could cause problems.

Despite the fact that the tradition of assigning numbers to the verbal categories on the 'words only' 9-point hedonic scale has lasted for approximately 60 years and has become accepted as a norm, it does not mean that it is a good idea. Such 'numbers' are merely alternative names for the categories and consequently are no more numerical than ranks. This creates obvious problems if they are treated as 'interval' data sampled from normally distributed populations and are subjected to parametric statistical analysis or modeling. Because the spacing between ranks is uncontrolled, normal distributions would be distorted rendering statistical analyses or models to be approximate. If users are aware of this, allowances can be made. Yet, problems of interpretation will occur for those users who are unaware of the true nature of their numbers. Similar arguments can be advanced for 'numbers only' hedonic scales where the cognitive strategy is to rank the products and use numerical values as an approximate measure of the spacing between the ranks.

The continued use of the serial monadic protocol for 'numbers only' scales can be called into question because it ignores the problem of forgetting and assumes the incorrect absolute model for the cognitive strategies used by consumers. Accordingly, based on the more correct relative model, the use of the rank-rating protocol would seem more appropriate. This protocol is simple and appropriate in cases where the products do not exhibit excessive adaptation effects which can hinder re-tasting. However, with the use of a 'words only' hedonic scale, where the goal is to measure liking per se rather than preference, it can be argued that the task is to match hedonic strengths to verbal exemplars. Accordingly, a serial monadic protocol can be argued to be appropriate. Yet, mistakes due to forgetting will still be encountered.

The serial monadic protocol would also be appropriate in the rare cases where trained judges had been calibrated for intensity like a calibrated laboratory instrument. Yet, calibration, even for simple estimations of stimulus concentration, requires several weeks of practice if the concentration exemplars are to be remembered accurately.⁸⁵ The Spectrum method of descriptive analysis⁸⁶ also uses a form of calibration for their panelists for judging attribute

strength. It uses a calibration to scale values rather physical quantities.

Although mean scores for the numbers assigned to the categories on a 'words only' hedonic scale are generally used, the representation of the data using histograms would seem preferable. A histogram clearly displays the distribution of responses to the various categories on the scale and any central tendencies. It should, at least, be used as an accompaniment to the traditional analysis. If a statistical analysis is mandatory, simple analyses based on binomial or chi-squared statistics can readily be adopted for comparison among products.

To avoid the potential problems with scaling, hedonic ranking with R-Index analyses is simple for consumers and produces data that are based on human behavior rather than consumers' sometimes questionable skills at using rating scales. Furthermore, such data (probabilities) are susceptible to parametric statistical analysis and do not suffer from the range compression encountered in East Asian countries. Also, for measures like product optimization where many products are assessed, a version of two-stage ranking based on initial categorization using the 'words only' version of the 9-point hedonic scale, can solve the problem of obtaining numerical measures of hedonic strength, when memory problems associated with too many products can distort the data. Despite its obvious advantages, it is anticipated that such analyses will be used as secondary to the traditional analysis for a while.

The purpose of this paper was merely to provoke some thoughts, stimulate some questions and maybe encourage some changes in the analysis or at least some additions to the analysis of 9-point hedonic scales.

REFERENCES

- 1 Lim J, Hedonic scaling: A review of methods and theory. *Food Qual Pref* **22**:733–747 (2011).
- 2 Rosas-Nextcapa M, Angulo O and O'Mahony M, How well does the 9-point hedonic scale predict purchase frequency? *J Sens Stud* **20**:313–331 (2005).
- 3 Lim J and Fujimaru T, Evaluation of the labelled hedonic scale under different experimental conditions. *Food Qual Pref* **21**:521–530 (2010).
- 4 Lim J, Wood A and Green BG, Derivation and evaluation of a labeled hedonic scale. *Chem Sens* **34**:739–751 (2009).
- 5 Cardello AV and Schutz HG, Numerical scale-point location for constructing the LAM (labeled affective magnitude) scale. *J Sens Stud* **19**:341–346 (2004).
- 6 Schutz HG and Cardello AV, A labeled affective magnitude (LAM) scale for assessing food liking/disliking. *J Sens Stud* **16**:117–159 (2001).
- 7 Schifferstein HNJ, Labeled magnitude scales: A critical review. *Food Qual Pref* **26**:151–158 (2012).
- 8 Peryam DR and Girardot NF, Advanced taste-test method. *Food Eng* **24**:58–61, 194 (1952).
- 9 Peryam DR and Pilgrim FJ, Hedonic scale method of measuring food preferences. *Food Technol* **11**:9–14 (1957).
- 10 Amerine MA, Pangborn RM and Roessler EB, *Principles of Sensory Evaluation of Food*. Academic Press, New York, pp. 367–374 (1965).
- 11 Kemp SE, Hollowood T and Hort J, *Sensory Evaluation: A Practical Handbook*. Wiley-Blackwell, Chichester, pp. 129–134 (2009).
- 12 Lawless HT and Heymann H, *Sensory Evaluation of Food. Principles and Practices*, 2nd edition. Springer, New York, pp. 326–329 (2010).
- 13 Peryam DR, Polemis BW, Kamen JM, Eindhoven J and Pilgrim FJ, *Food Preferences of Men in the U. S. Armed Forces*. Department of the Army Quartermaster Research and Engineering Command-Quartermaster Food and Container Institute for the Armed Forces, Chicago, IL (1960).
- 14 Stone H and Sidel JL, *Sensory Evaluation Practices*, 3rd edition. Elsevier Academic Press, London, pp. 87–90 (2004).
- 15 Zwislocki JJ, Absolute and other scales: Question of validity. *Percept Psychophys* **33**:593–594 (1983).
- 16 Zwislocki JJ and Goodman DA, Absolute scaling of sensory magnitudes: A validation. *Percept Psychophys* **28**:28–38 (1980).

- 17 Mellers BA, Evidence against "absolute" scaling. *Percept Psychophys* **33**:523–526 (1983).
- 18 Mellers BA, Reply to Zwolsick's views on "absolute" scaling. *Percept Psychophys* **34**:405–408 (1983).
- 19 Parducci A, Range–frequency compromise in judgment. *Psychol Monogr* **77**:1–50 (1963).
- 20 Parducci A, Category judgment: A range–frequency model. *Psychol Rev* **72**:407–418 (1965).
- 21 Parducci A, The relativism of absolute judgments. *Sci Am* **219**:84–90 (1968).
- 22 Schifferstein HNJ and Frijters JER, Contextual and sequential effects on judgments of sweetness intensity. *Percept Psychophys* **52**:243–255 (1992).
- 23 Stillman JA, Context effects in judging taste intensity: A comparison of variable line and category rating methods. *Percept Psychophys* **54**:477–484 (1993).
- 24 Lawless HT and Malone GJ, A comparison of rating scales: Sensitivity, replicates and relative measurement. *J Sens Stud* **1**:155–174 (1986).
- 25 Poulton EC, Models for biases in judging sensory magnitudes. *Psychol Bull* **86**:777–803 (1979).
- 26 Nicolas L, Marquilly C and O'Mahony M, The 9-point hedonic scale: Are words and numbers compatible? *Food Qual Pref* **21**:1008–1015 (2010).
- 27 O'Mahony M, De Mingo N, Mendez Martin J, Perret M and Bodeau J, 9-point hedonic scales: data derived from versions using words or numbers give different results, while serial monadic presentation can reduce power, in *Abstracts of 9th Pangborn Sensory Science Symposium, 4–8 September, Toronto, Canada*. Elsevier, Oxford, UK, Abstract # P1.8.37 (2011).
- 28 Kim K-O and O'Mahony M, A new approach to category scales of intensity I: Traditional vs rank-rating. *J Sens Stud* **13**:241–249 (1998).
- 29 Colyar JM, Eggett DL, Steele FM, Dunn ML and Ogden LV, Sensitivity comparisons of sequential monadic and side-by-side presentation protocols in affective consumer testing. *J Food Sci* **74**:S322–S327 (2009).
- 30 Jeon S-Y, O'Mahony M and Kim K-O, A comparison of category and line scales under various experimental protocols. *J Sens Stud* **19**:49–66 (2004).
- 31 Koo T-Y, Kim K-O and O'Mahony M, Effects of forgetting on performance on various intensity scaling protocols: Magnitude estimation and labeled magnitude scale (Green scale). *J Sens Stud* **17**:177–192 (2002).
- 32 Lee H-J, Kim K-O and O'Mahony M, Effects of forgetting on various protocols for category and line scales of intensity. *J Sens Stud* **16**:327–342 (2001).
- 33 Stevens SS, Mathematics, measurement, and psychophysics, in *Handbook of Experimental Psychology*, ed. by Stevens SS. John Wiley, New York, pp. 1–49 (1951).
- 34 Stevens SS, The psychophysics of sensory function. *Am Sci* **48**:226–253 (1960).
- 35 Kim K-O, Ennis DM and O'Mahony M, A new approach to category scales of intensity II: Use of d' values. *J Sens Stud* **13**:251–267 (1998).
- 36 Dorfman DD and Alf E, Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals – rating-method data. *J Math Psychol* **6**:487–497 (1969).
- 37 Ogilvie JC and Creelman CD, Maximum-likelihood estimation of receiver operating characteristic curve parameters. *J Math Psychol* **5**:377–391 (1968).
- 38 Hautus M, O'Mahony M and Lee H-S, Decision strategies determined from the shape of the same–different ROC curve: What are the effects of incorrect assumptions? *J Sens Stud* **23**:743–764 (2008).
- 39 O'Mahony M and Hautus M, The signal detection theory ROC curve: Some applications in food sensory science. *J Sens Stud* **23**:186–204 (2008).
- 40 Wichchukit S and O'Mahony M, A transfer of technology from engineering: use of ROC curves from signal detection theory to investigate information processing in the brain during sensory difference testing. *J Food Sci* **75**:R183–R193 (2010).
- 41 Green DM and Swets JA, *Signal Detection Theory and Psychophysics*. Robert E. Krieger, New York, pp. 87–91 (1974).
- 42 Macmillan NA and Creelman CD, *Detection Theory: A User's Guide*, 2nd edition. Lawrence Erlbaum Associates, London, pp. 31–34 (2005).
- 43 Kim H-J, Kim K-O, Jeon S-Y, Kim J-M and O'Mahony M, Thurstonian models and variance II: experimental confirmation of the effects of variance on Thurstonian models of scaling. *J Sens Stud* **21**:485–504 (2006).
- 44 Jones LV and Thurstone LL, The psychophysics of semantics: an experimental investigation. *J Appl Psychol* **39**:31–36 (1955).
- 45 Jones LV, Peryam DR and Thurstone LL, Development of a scale for measuring soldiers' food preferences. *Food Res* **20**:512–520 (1955).
- 46 Edwards AL, The scaling of stimuli by the method of successive intervals. *J Appl Psychol* **36**:118–122 (1952).
- 47 Stroh S, Response bias and memory effects for selected scaling and discrimination protocols. MS thesis, University of California, Davis (1998).
- 48 Oppenheim AN, *Questionnaire Design, Interviewing and Attitude Measurement*. Pinter, London (1992).
- 49 Allgeyer LC, Miller MJ and Lee S-Y, Drivers of liking for yogurt drinks with prebiotics and probiotics. *J Food Sci* **75**:S212–S219 (2010).
- 50 Hong JH, Yoon EK, Chung SJ, Chung L, Cha SM, O'Mahony M, et al., Sensory characteristics and cross-cultural consumer acceptability of *Bulgogi* (Korean traditional barbecued beef). *J Food Sci* **76**:S306–S313 (2011).
- 51 Yeu K, Lee Y and Lee S-Y, Consumer acceptance of an extruded soy-based high-protein breakfast cereal. *J Food Sci* **73**:S20–S25 (2008).
- 52 Illupapalayam VV, Smith SC and Gamlath S, Consumer acceptability and antioxidant potential of probiotic-yogurt with spices. *LWT – Food Sci Technol* **55**:255–262 (2014).
- 53 Bindon K, Holt H, Williamson PO, Varela C, Herderich M and Francis IL, Relationships between harvest time and wine composition in *Vitis vinifera* L. cv. Cabernet Sauvignon 2. Wine sensory properties and consumer preference. *Food Chem* **154**:90–101 (2014).
- 54 Fan X and Sokorai KJB, Changes in quality, liking, and purchase intent of irradiated fresh-cut spinach during storage. *J Food Sci* **76**:S363–S367 (2011).
- 55 Hein KA, Jaeger SR, Carr BT and Delahunty CM, Comparison of five common acceptance and preference methods. *Food Qual Pref* **19**:651–661 (2008).
- 56 Lovely C and Meullenet J-F, Comparison of preference mapping techniques for the optimization of strawberry yogurt. *J Sens Stud* **24**:457–478 (2009).
- 57 Ngamchuachit P, Sivertsen HK, Mitcham EJ and Barrett DM, Effectiveness of calcium chloride and calcium lactate on maintenance of textural and sensory qualities of fresh-cut mangos. *J Food Sci* **79**:C786–C794 (2014).
- 58 Hayes JE, Raines CR, DePasquale DA and Cutter CN, Consumer acceptability of high hydrostatic pressure (HHP)-treated ground beef patties. *LWT – Food Sci Technol* **56**:207–210 (2014).
- 59 Bonany J, Brugger C, Buehler A, Carbo J, Codarin S, Donati F, et al., Preference mapping of apple varieties in Europe. *Food Qual Pref* **32**:317–329 (2014).
- 60 Cheng VJ, Bekhit AE-D, Sedcole R and Hamid N, The impact of grape skin bioactive functionality information on the acceptability of tea infusions made from wine by-products. *J Food Sci* **75**:S167–S172 (2010).
- 61 Lavelli V, Sri Harsha PSC, Torri L and Zeppa G, Use of winemaking by-products as an ingredient for tomato puree: The effect of particle size on product quality. *Food Chem* **152**:162–168 (2014).
- 62 Lebesi DM and Tzia C, Staling of cereal brand enriched cakes and the effect of an endoxylanase enzyme on the physicochemical and sensorial characteristics. *J Food Sci* **76**:S380–S387 (2011).
- 63 Chung S-J, Effects of milk type and consumer factors on the acceptance of milk among Korean female consumers. *J Food Sci* **74**:S286–S295 (2009).
- 64 Lee J and Chambers DH, Descriptive analysis and U.S. consumer acceptability of 6 green tea samples from China, Japan, and Korea. *J Food Sci* **75**:S141–S147 (2010).
- 65 Lee J, Chambers IV E, Chambers DH, Chun SS, Oupadissakoon C and Johnson DE, Consumer acceptance for green tea by consumers in the United States, Korea and Thailand. *J Sens Stud* **25**(Suppl S1):109–132 (2010).
- 66 Brown J, Recognition assessed by rating and ranking. *Br J Psychol* **65**:13–22 (1974).
- 67 O'Mahony M, Understanding discrimination tests: A user-friendly treatment of response bias, rating and ranking R-Index tests and their relationship to signal detection. *J Sens Stud* **7**:1–47 (1992).
- 68 Lee H-S and van Hout D, Quantification of sensory and food quality: the R-Index analysis. *J Food Sci* **74**:R57–R64 (2009).
- 69 Bi J and O'Mahony M, Updated and expanded table for testing the significance of the R-Index. *J Sens Stud* **22**:713–720 (2007).

- 70 Argai A, Pérez-Vega O and López-Malo A, Sensory detection of cooked flavor development during pasteurization of a Guava beverage using R-Index. *J Food Sci* **70**:S149–S152 (2005).
- 71 Kappes SM, Schmidt SJ and Lee S-Y, Mouth-feel detection threshold and instrumental viscosity of sucrose and high fructose corn syrup solutions. *J Food Sci* **71**:S597–S602 (2006).
- 72 Robinson KM, Klein BP and Lee S-Y, Utilizing the R-Index measure for threshold testing in model caffeine solutions. *Food Qual Pref* **16**:283–289 (2005).
- 73 Chauhan J and O'Mahony M, Use of a signal detection ranking analysis to measure preference for commercial and 'health modified' cakes. *J Sens Stud* **8**:69–75 (1993).
- 74 Pasin G, O'Mahony M, York G, Weitzel B, Gabriel L and Zeidler G, Replacement of sodium chloride by modified potassium chloride (cocrystallized disodium-5'-inosinate and disodium-5'-guanylate with potassium chloride) in fresh pork sausages: Acceptability testing using signal detection measures. *J Food Sci* **54**:553–555 (1989).
- 75 Vié A, Gulli D and O'Mahony M, Alternative hedonic measures. *J Food Sci* **56**:1–5, 46 (1991).
- 76 Pipatsattayanuwong S, Lee H-S, Lau S and O'Mahony M, Hedonic R-Index measurement of temperature preference for drinking black coffee. *J Sens Stud* **16**:517–536 (2001).
- 77 Lee H-S and O'Mahony M, Sensory evaluation and marketing: Measurement of a consumer concept. *Food Qual Pref* **16**:227–235 (2005).
- 78 Lee H-S and O'Mahony M, Some new approaches to consumer acceptance measurement as a guide to marketing. *Food Sci Biotechnol* **16**:863–867 (2007).
- 79 Simone M and Pangborn RM, Consumer acceptance methodology: one vs. two samples. *Food Technol* **6**:25–29 (1957).
- 80 Villegas-Ruiz X, Angulo O and O'Mahony M, Hidden and false "preferences" on the structured 9-point hedonic scale. *J Sens Stud* **23**:780–790 (2008).
- 81 Yeh LL, Kim K-O, Chompreeda P, Rimkeeree H, Yau NJN and Lundahl DS, Comparison in use of the 9-point hedonic scale between Americans, Chinese, Koreans, and Thai. *Food Qual Pref* **9**:413–419 (1998).
- 82 Yao E, Lim J, Tamaki K, Ishii R, Kim K-O and O'Mahony M, Structured and unstructured 9-point hedonic scales: A cross cultural study with American, Japanese and Korean consumers. *J Sens Stud* **18**:115–139 (2003).
- 83 Feng Y, A critical appraisal: the 9-point hedonic scale and hedonic ranking as an alternative. MS thesis, University of California, Davis (2012).
- 84 Birch LL, Birch D, Marlin DW and Kramer L, Effects of instrumental consumption on children's food preference. *Appetite* **3**:125–134 (1982).
- 85 O'Mahony M and Wong S-Y, Time-intensity scaling with judges trained to use a calibrated scale: Adaptation, salty and umami tastes. *J Sens Stud* **3**:217–236 (1989).
- 86 Meilgard M, Civille GV and Carr BT, *Sensory Evaluation Techniques*, 2nd edition. CRC Press, Boca Raton, pp. 173–185 (1991).