Scientific measurement of sensory preferences using stimulus tetrads

Article (Accepted Version)

Journal of Sensory Studies accepted for publication on 2 March 2012

Running title: SENSORY PREFERENCE TETRADS

SCIENTIFIC MEASUREMENT OF SENSORY PREFERENCES USING STIMULUS TETRADS

DAVID A. BOOTH[1]

Food Quality Research Group, School of Psychology, College of Life and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, U.K.

[1]Corresponding author.
EMAIL: D.A.Booth *at* Bham.ac.UK *and* D.A.Booth *at* sussex.ac.uk

ABSTRACT <151 words

This paper provides the evidence base to construct a professional standard for discriminative scaling of taints and optima. The measurement of suboptimal sensed characteristics of a product has logical and empirical requirements that specify a single overall rating of each sample in a tetrad. Those four pairs of response/stimulus data determine the discrimination distance of each sample from the comparison in memory used by the assessor, together with the position of that standard on the straight line specified by the two stimulus levels in the tetrad. The rating's reference anchor can be the match to a familiar version of the product or the personally most preferred level. Each sample can be assessed again for sensory and/or conceptual attributes, using vocabulary learnt in life or by sensory training. Those data give the ideal or matching value of that verbal category, and the individual's tolerance of deviations from that value. 148 words

PRACTICAL APPLICATIONS <151 words

This paper is solely concerned with the practical use of difference tests to measure sensory taints and, by extension, to carry out to sensory optimisation with considerable economy. Individual consumers' discriminatively scaled ideal or matching points can be aggregated across a purpose selected panel to give profiles of estimated market response to sensory intensities, concept impacts, and/or product constituent levels, depending on the data collected. This scientifically fundamental approach to practical issues faced by sensory studies gives objective results that are specific, precise and directly relevant to the detection and identification of taints and the optimisation of formulations for the market or the segment sampled by selected panel. 108 words

Keywords up to 6
Sensory Preferences. Taint Measurement. Optimisation. Sensory Characterisation. Sensory Impact. 5

INTRODUCTION

This paper provides the theory and evidence that could form the basis for a new sensory standard to measure differences between food products or other familiar materials. Instead of the usual statistical models of grouped data, this approach provides scientific measurements of the performance achieved by each practised user of a material, at no extra cost of time or formulations. During sensory testing, as in life, the individual assessor discriminates between each product sample and a reference in memory of that product's past uses. These disparities between test samples and the familiar or ideal version can be measured by ratings of preference or choice, with or without additional ratings using sensory vocabulary. That refutes the continuing supposition that behavioural responses must be gathered separately from descriptive analyses. The precision and immediate business relevance of these measurements has been shown many times since the 1980s, as cited at relevant points in this practical exposition.

Discrimination Tetrads

How well an individual assessor distinguishes between levels of a sensed factor can be calculated from one rating of sensory strength and/or preference for each of the four blind samples of a 'tetrad', i.e. duplicates of the two differently sourced materials. These four pairs of data provide three distinct measurements: (i) that assessor's discriminative acuity for the sensed overall difference between the materials, traditionally called the 'just-noticeable difference' or, more objectively, the Weber fraction or half-discriminated disparity; (ii) the optimum level of the tested constituent(s), whether a normal part of the familiar product or a tainting foreign compound (or mixture); (iii) the size of the sensed overall difference between the two materials in units of discriminative performance. In the case of preference ratings, this third measure is the assessor's behavioral tolerance of that deviation from optimum.

Those estimates can be made more precise by rating the four samples for a second time in the same session, if there is no satiation. The later part of the test session can include sensory concepts, specified in panel-trained vocabulary or, more sensitively, in free choice of terminology learned in life (Williams and Arnold 1985, Booth 2014). The same principles apply to non-sensory conceptual factors in preference, whether symbolised on the test sample (Booth and Freeman 1993, 2014) or brought to the unlabeled sample by the assessor (e.g., Chechlacz and others 2009; cp. Thomson and Crocker 2014).

Logical requirements for identifying a sensory or conceptual factor in preference are described first below. Then the paper summarises empirical constraints from psychological theory on the measurement of sensory impact on preference. This approach is briefly compared with established procedures and outcomes of difference tests. Examples are given of the treatment of raw data that yields the objective sensed distance between the two materials and a distribution across the panel of the discriminative performance by each individual. If the panel represents users of the product, that aggregate is a direct estimate of responses from the market to the variants tested in a simulation of the use situation.

LOGIC OF CAUSATION

Measurement of a causal process such as the influence of a sensory (or conceptual) factor on acceptance of a product has three necessities of logic, ahead of any specifically scientific or statistical issues. 1. Variations in the factor under investigation must not be confounded with variations in other potential influences on preference. 2. At least two levels of the hypothesised factor must be tested. 3. The sensory effect on acceptance must be observed, not just experienced subjectively.

These three pre-conditions for successful measurement of any mechanism may seem obvious. Nevertheless they are commonly breached in sensory studies.

The three requirements were met in an interlaboratory study of sensory influences on preferences among brands of milk chocolate popular in the UK (Conner, Booth and Haddon 1987). Each consumer provided sensory descriptors for the first two or three samples with brand identification removed. All the brands were then rated blind for sensory intensity relative to the personally most preferred level (ideal point) using the panelist's own vocabulary. Mean unfolded sensory scores from the panel of consumers served as surrogates for the sucrose, milk powder and cocoa butter contents that varied in proportions among the seven brands (e.g. Figure 3 in Booth 1988b). Multiple regression from folded sensory scores onto overall preferences within each panelist characterised four distinct sensory segments. The modal preference ranks collected by the study leader were attributable to a minority segment in our panel being in the majority in the leader's panel. Statistically more elaborate approaches by other laboratories were unable to account for the leader's observations.

Unconfounded Influences

Orthogonal rotation of the sensory responses was sufficient to establish a degree of independent variation of constituents across the brands tested (logic 1). Note that in this case there was no need to prepare any experimental variants because there were sufficiently low correlations between products on the market.

The effects of two putative causes cannot be distinguished unless the variations in those factors are minimally correlated. Complete confounding makes it logically impossible to pick out the effects of one factor from the other. In practice, a modest correlation between the levels of two influences may not prevent separation of their effects on preference or familiarity (Booth and others 2003). Even when $r = 0.7$, half the variance remains available for distinguishing between the effects.

Two-dimensional plots of sensory profiling vocabulary from principal components analysis, multidimensional scaling of preferences, and other approaches, often show two or more descriptors on or close to the same vector. Such confounding is readily avoided by classic

statistical procedures for data-reduction, namely orthogonal factor analysis followed by linear combination of the variance in identified factors using multiple regression. At the group level, multivariate analysis of variance achieves the same disconfounding in a single step. However, the two-step approach can be applied individual by individual.

Two Levels

The factor analysis also demonstrated that there were at least two levels of each sensory component (logic 2). Nevertheless, just two levels of a putative factor are sufficient to measure the strength of that influence. Furthermore, least squares regression can be calculated from as few as three observations, e.g., a rating of each sample in the odd-one-out stimulus design used in the triangle test. At one of the two levels, however, such a triad provides no basis for an estimate of the variance in responses. Hence a tetrad of samples is preferable, purely from the logic of replication.

Regression through the observed response/stimulus pairs provides an estimate of the discriminative acuity of an individual assessor's or a whole panel's sensory judgments (Torgerson 1958). The finer the acuity, the stronger is the influence of that stimulus on the response. How finely the response discriminates between levels of the stimulus is simply the mathematical inverse of the causal power of the stimulus over the response.

The Size of the Difference

When observed in different responses to different stimuli (logic 3), the sensed amount of the disparity between samples of two materials is a matter of fact about human achievement. Logically, therefore, a measurement of a physical or chemical difference between the materials is not sufficient. Neither is a difference merely between the assessors' quantitative responses to the samples. The size of a sensory difference is a parameter of the causal relation within the assessor's mind between the socially meaningful response quantities and the physicochemical stimulus quantities. The scientific issue for a difference test is how to measure the power of the impact of a sensed difference on a person's performance on the pair of materials.

The discrimination may be achieved in conceptual terms or by action. That is to say, the response can be either judgments of the intensity of a verbally labelled sensory characteristic, or quantitative expressions of the disposition to accept the sample for a particular use. Both were obtained for the chocolates in the interlaboratory tests (above).

The size of a sensed difference between materials should not be confused with the reliability of an estimate of the size of that difference. Other things being equal, it is the precision of a measure, not its numerical value, which increases with the number of samples measured. Hence an assessor's discriminative acuity does not necessarily improve when the same testing procedure is moved from a triad to a tetrad of samples. The greater measurement

power of tetrads than triads is a logical implication from the amounts of data, not a factual issue about the sensory difference being tested for.

The basic statistical measure of the reliability of an effect is the proportion of variance accounted for, as in the square of a regression coefficient or of the partial eta from ANOVA. The reliability of a particular estimate should be determined from its confidence limits in the units of measurement, not from a low probability that the true estimate is zero (MacRae 1995). The number of panellists needed to get a *P* value less than the conventional 5% is not a measure of how widely two materials are perceived to differ, nor of the importance of the effect of any difference on the activities of the material's users. The size of a difference and its behavioral significance are psychological measurements -- of each assessor too, not of the panel in the first instance.

Sequence of Test Samples

A major concern about any measurement of the performance of an adapting system, such as a human mind, is an effect of an earlier procedure on performance at a later repetition. Indeed, it is conceivable that the sheer passage of time could affect the measurement operation or the performance characteristics of the system itself.

Any such biases from the passage of time, repeated presentations, or the sequencing of the two types of sample (A and B), may largely be balanced out across a panel by pseudo-random assignments of the sequences ABBA and BAAB. Either sequence also counters most such biases within the individual. Hence sessions of monadic tests of stimulus tetrads are best carried out in these counterbalanced sequences. If a full examination of time or sequence effects in tetrads is required, then the contrast between ABAB and BABA should be included in addition.

REQUIREMENTS FROM EMPIRICAL THEORY

Weber's Fraction

The most fundamental general principle of the sensing of differences between materials was discovered by E.H. Weber in the early 1840s (Ross and Murray 1996). Weber himself (1834/1996) and many others subsequently showed that the principle applied across sensory modalities. When the stimulation is moderate, the minimum detected disparity between two levels of a stimulus is a constant fraction of the physical measure of those stimuli. In other words, discrimination in such a range is achieved at a constant ratio of physical units, which is a constant interval in a logarithmic conversion. That is to say, the logarithm of the physical quantity of a tested stimulus plotted against the quantitative value of the response to each of those stimuli forms a straight line (Fechner 1860/1966).

Unfortunately, Fechner ignored a major feature of Weber's evidence. That neglect was perpetuated by some subsequent theorists of mental scaling, such as Thurstone (1927) and Stevens (1961). The Weber fraction increases at both low and high levels of stimulation (as shown in textbook presentations of typical findings). That is, the semi-logarithmic plot of observed responses to presented physical stimuli is a straight line in the middle but bends over at the extremes: the slope steadily decreases towards either undetectably low levels or receptor-saturating high levels. In terms of the assessor's performance, the response becomes less and less sensitive to disparities in strength of stimulation.

Hence Fechner was wrong to extend the linear semi-logarithmic function to the zero response. To obtain linearity of responding, the physical stimulus levels must remain with the region of constancy of Weber's fraction. The proper way to accommodate the facts is to construct an empirical theory that places the zero point of the semi-logarithmic plot somewhere in that medium range of stimulation. The present approach is to use the familiar or most preferred level as the zero, and also as the primary anchor on the layout for making a response to each sample. As a result, this approach both measures the basic mechanisms of sensory appreciation and also provides exact answers to practical questions about failures of products to match familiar or ideal formulations.

It should be noted that this use of the observed facts has nothing to do with any version of the Weber-Fechner "Law", which was intended to link the consciousness of sensations to the material universe. How an assessor manages to keep the Weber fraction constant is a scientific issue that can only be resolved with more evidence than only one varied stimulus and one quantitative response (Booth and Freeman 1993).

The Constant Comparison in Memory

Another neglected aspect of a range-limited Weber-Fechner graph implies exactly where its zero point is. Each plotted stimulus value has its own response value. The graph does not include any response specific to a constant second stimulus presented alongside. Indeed, the equation for the linear region of the semilogarithmic plot specifies by itself a level which is not physically presented but by which each test stimulus sample was evaluated when the response was generated. This implicit standard is traditionally known as the point of equality (Torgerson 1958). There is no need for judgments relative to a comparison stimulus, as in Weber's original experiments, Thurstone's procedures for comparing each test sample with a standard sample, and Stevens's use of a modulus sample as one anchor for numerical ratings of test stimuli. Monadic judgments can and should be elicited and plotted.

The stimulus level in the implicit comparison standard is not necessarily the same as the level in any presented standard. The point of equality calculated from the data is a physical level held in long-term memory and brought by the assessor to the judgment on each test sample. Such previously learned personal standards, adapted if necessary to the experimental stimuli, provide more precise results than the conventional external standards (e.g., Morgan and others 2000, Nachmias 2006). If the test samples and testing situation are too artificial to be

comparable with any previous experience, then an internal standard is constructed from the initial test samples (Stewart and others 2005).

The point of equality to the most relevant familiar level is the zero point that Fechner and his followers should have used. This personal and contextualised ('situated') norm in memory should replace Thurstone's standard of comparison and Stevens's modulus in sensory studies. Assessment needs to be monadic -- presenting samples one at a time and eliciting a norm-relative response to each sample by itself. Indeed, when samples are presented in dyads, each stimulus is automatically compared with the norm in memory; then those two disparities from norm have to be compared in short-term memory in order to construct a comparative response to the dyad. Direct judgment of each sample relative to a learned norm is likely to be both more accurate and also easier to carry out.

These judgments of the quantity of stimulation can be directed to the particular standard that is of interest to the investigator within the situation that is simulated by the testing procedures. If perception of a familiar brand is to be compared with personal preference, then two responses can be elicited, one anchored on the name of the branded product and the other on the concept of the ideal.

In accordance with this theory, the rating of each sample, i.e. monadic testing, was introduced for both or either the degree of preference (strength of disposition to accept) and/or the strength of stimulation (Booth 1988a, Booth and others 1983, Conner and others 1986). New analyses of the earliest data of this sort (Booth and others 1983) are used here to illustrate discriminative difference measurement using tetrads (and triads) of stimulus samples.

TASK INSTRUCTIONS AND RATING FORMATS

Laboratory investigators attempt to constrain human participants to particular tasks by verbal communication, oral and/or written. In sensory tests, the assessor is asked to use a specified procedure to examine each sample, or dyad or larger set of samples, and to make responses in a particular way. Obviously the instructions, stimulus presentations and response layouts should be self-consistent and clearly so. What may be less obvious is that all of this should also be consistent with scientific theory of human performance in situations like those of the test. The feasibility of statistical analysis of the data is insufficient.

Scales, Scales and Scales

The term 'scale' has been given three radically different meanings in areas relating to applied sensory research (Booth 2009). First, the objective psychological scale of mental performance is the linear relationship achieved between response values and stimulus values. Second, the axes of a graph are called scales; in sensory research, these axes may be of material or conceptual quantities. Third, the way in which responses are assigned to locations on a number line is commonly called a scale; the assessor's positioning of the mark may be

on a continuous line, a row of boxes, a series of numbers, or even an ordered set of phrases. Expositions of sensory scaling often slide between these totally different meanings of the term.

A major culprit for confusion of a scale in the mind with physically manifest answers to questions is the introspectionist assumption that Fechner and his followers have imposed on the response-stimulus function. Contrary to that doctrine, one type of response by itself cannot measure the strength of a privately experienced sensation. All that is observed is the quantitative use of a socially agreed concept. There may be no stimulus-specific subjective magnitude accompanying the influence of different levels of stimulation on the degree of preference, when that is measured without bringing the relevant sensory concept to mind (Booth and others 2011b). Mere use of sensory vocabulary should not be called perception, or description either, unless the evidence from those data shows that those conceptual quantities are strongly influenced by the varied stimulation, and not by any other stimulation.

Monadic Ratings

To provide a measure of a perceived sensory difference, the assessor needs only to place each sample at a location on a number line that represents physical strengths of stimulation. In a graph of the data, the two materials in a tetrad or triad are at points on the stimulus axis, conventionally the abscissa (x axis). The responses are at points on the ordinate (y axis), yielding a y/x function of responses on stimuli.

The two points on the stimulus axis may be dummy coded, e.g. as zero for the material nearer to the familiar or ideal variant and any finite number for the other material. A code of 1 is convenient mathematically and also follows the convention of 1 and 0 standing for presence and absence (of proximity to the norm in this case). If a physical measure of the difference is available (even of only one component of a mixture in fixed ratios), these two tested physical values can be plotted on the x axis after logarithmic transformation of the units in accord with the constancy of discrimination ratios that was discovered by Weber (McBride 1983).

On the response axis, the points also have to be locations on a number line. A continuous line is not an analogue of anything psychological. It merely represents a segment of the continuum of real numbers. A regularly broken line or a row of boxes represents part of the series of small integers. Hence any linear display can be used to make quantitative judgments about something by indicating points on the line. What those judgments are about is determined by the judge's assignments of the numbers represented to the variation in the tested samples.

Any straight line is specified by just two points. Hence, contrary to S.S. Stevens's recommendation of a single modulus sample, there has to be more than one reference anchor. Contrary also to the widespread assumption that people are capable only of using a row of boxes if there is a word, phrase or sentence alongside each box, there cannot be more than two such quantitative anchor categories of a single concept. If there are three or more anchor

phrases, each adjacent pair is liable to generate a y/x function having a different slope from the other anchor pairs.

There is one special case where three reference points might be usable. If the mid-category is the personal ideal or the target-matching point, then the phrases “just too little” and “just too much” can be provided at the same distance on either side, together with room for more extreme responses (unlike the usual end-anchored line ratings). If the anchors of insufficiency and excess do initially differ in slope from the norm point, assessors force their use of those reference points to the same slope (Conner and Booth 1992). This is not possible with several diversely worded ordinal categories, as repeatedly shown by comparisons with numerical ratings (so-called ‘magnitude estimation’) and more definitively by Thurstone scaling (Jones and others 1955). The slopes that an individual assessor puts between quantitative categories in a particular experiment are not determinable until after those data have been collected. The rating format cannot reliably be rigged in advance.

Another dire error with multiple categories is the use of phrases that refer to ranges of points, rather than to a single point. This is especially dangerous when preferences are being assessed, not just intensities. The most disastrous example is conversion of the “just right” position on a line to a “just about right” (JAR) range at a check box. Quantitative interpretations of such data can put both companies and consumers at risk (Booth and Conner 2009). The individual’s ideal level of sweetener could be anywhere in a range with unspecified ends, whether or not unlabeled boxes are added (López Osornio and Gough 2010). When JAR formats are limited to only two outer anchors, e.g. “too little” and “too much”, the analysis of data has to be restricted to panellists who happen to have been tested on samples including one JAR; furthermore, range bias is corrected subjectively, and only a rough estimate can be made of a most popular level across the residue of the panel (Garitta and others 2006).

Almost as badly flawed is the opposite extreme of using an unmarked continuous line with an anchor point at each end (the misnamed ‘visual analog scale’). The task is to position responses on a dimension and so the quantitative judgments are better supported by verbally unlabeled marks, or short breaks in the line, or indeed a linear array of boxes or cups. If the judgments concern solely the strength of stimulation, then the lower anchor can be “none at all” at one end of the row of boxes or the structured line. An upper anchor of “extremely” is inadequate. Even “as strong/bad as imaginable” may not accommodate all samples. If stimulation happens to be stronger even than the assessor imagined at the start of the session, that may force rescaling, i.e. the y/x slope for subsequent samples will be flatter than that for previous samples.

Probably the best response to ask for is selection of one of a row of single-digit integers. So long as zero is included in that series, four or more low integers are likely to be as good as an unbroken straight line at showing interval and ratio properties among positions (Bowman and others 2004; Booth and others 2011a; Goodchild and others 2008). The numbers up to ten have conceptually meaningful quantitative uses that are very well practised by members of

the public, for example when assessing the performance of participants in televised competitions.

For validity and precision, each of the two anchors for the number line (quantity analog) needs to refer to a level with which the assessor is well acquainted. Hence one anchor should be the point matching the personal ideal or a familiar example, e.g. “just right” or “always choose”. Probably the most easily judged lower anchor is the limit of acceptance, where a judgment of unacceptable becomes equally likely. “Just too little” or “just too much” should not be used as an end anchor or as the label for zero, because it is logically possible that the sensed level could be worse than just tolerable. Nevertheless, such a possibility can be precluded in practise if there are any options among samples to present to an assessor. After two acceptable levels have been tested, a minimal response/stimulus function can be extrapolated through each of the anchors and the subsequent samples all kept at levels within that tolerated range (Conner and others 1986).

DISCRIMINATIVE SCALING

The Difference Triangle

A sensory difference could make material A preferred over material B. This greater acceptance of A could arise from its level of a sensory factor being higher or lower than the level in B. In other words, the degree of preference may be reduced by either too little or too much stimulation by a sensed characteristic. This principle of ‘folding’ of response levels on stimulus levels has long been recognised for sensory preferences (Coombs 1964). It was incorporated from the start into psychologists’ non-metric multidimensional modeling programs for influences on preference (Carroll 1972), much used lately in sensory studies.

Less widely recognised, closeness to familiar also comes from one of two directions -- unusually little or much of a feature. Since comparison with an acquired norm is implicit in each intensity rating, the ratings from quantitative sensory analysis (not just preference ratings) should in theory also be folded on the assessor’s standard level of the sensed variable.

The relation between response and stimulus quantities is linear over the region of constancy in Weber’s fraction. The semi-logarithmic plot from too little through the ideal point or best match towards too much is indeed fitted well by linear regression (e.g. Conner and Booth 1992). Folding at the ‘just right’ point therefore means that the response-stimulus function has the shape of an isosceles triangle, with the same numerical value of slope on either side of the apex but opposite signs (Booth and others 1983; Conner and Booth 1991; Conner, Booth, Clifton and Griffiths 1988).

Contextual Defects

In every sensory test, each factor under investigation is in a context of other sensed factors. Even the academic psychophysicists' pure solution of a single compound has at least temperature, viscosity and aspects of visual appearance, as well as taste. The basic scientific theory is that quantitative judgments on sets of artificial stimuli are achieved by assimilation to, contrast with or construction into a multifactor norm for a situation or task that has repeatedly occurred in the assessor's past (Booth and Freeman 1993).

In the assessment of a sensory factor for preference or intensity, other factors in the samples or the ambience of testing may be at some distance from their familiar or ideal levels. When that is the case, then overall preference or familiarity cannot be perfect, even when the investigated sensory factor is at the right level. In other words, the sharp apex of the theoretical acceptance triangle is not observed if a contextual factor departs substantially from the norm on average during a session. Blunted peaks are indeed widely observed in individuals' data. (These rounded curves should not be confused with the smooth and wide peaks seen in average preference scores or counts of ideal points across a panel: that rounding arises from variation across individuals in the matching points, however sharply pointed each person's y/x function is.)

Hence, the rounding of an individual's peak of preference for a sensory factor is evidence that the session of testing was generally a poor simulation of the context in which the varied stimulus is usually sensed. One or more of the other sensory factors may have been seriously 'off', or the attribution of some concept has degraded the assessor's opinion of all the samples. The extent to which the peak is lowered from the apex of the difference triangle reflects the degree of defectiveness in the context, even when the physical basis of the defect is unknown. If such data are unfolded, a discontinuity is generated at the ideal point (Booth 1994). As a result, there is greater variance in that region in a linear regression (Conner, Haddon and Booth 1986). There is no evidence that the individual ceases to distinguish or be motivated by differences in level, as claimed by Garitta and others (2006). Rather, the defect is too great for those samples to approach closer to the overall ideal.

Variations in any contextual feature, or indeed in the contextual configuration as a whole, can be represented in another isosceles triangle at a right angle to the sensory factor that is being varied systematically. If levels of the focal sensory variable are plotted on the x axis, with response levels on the y axis, the contextual triangle can be plotted on the z axis, going through the x,y plane from front to back. The resulting three-dimensional layout of the data from a sensory test has the shape of a cone. When the mean of the contextual levels is to one side of the apex, the data will fall on the surface of a vertical cut though that 'side' of the cone. These conic sections have rounded peaks. The difference between the observed rounded peak and the theoretical triangle's apex, in either the response score or the stimulus level, provides a measure of the defect in context across all the factors that have an impact (Booth & Freeman 1993; invited presentation at the first Pangborn Symposium, Helsinki 1992).

Hyperbolic Folding

The equation for a perpendicular conic section is a hyperbola, $y^2/x^2 = k$. Therefore all sensory data should be fitted to this equation before further analysis. This procedure contrasts with the statistical tradition of fitting polynomials to data. That approach specifies a quadratic equation for the folding of monotonic data into the peaked sensory preference function that is observed in accord with psychological theory (Coombs 1964, Carroll and Chang 1970). A hyperbola was initially proposed for the weaker response to deviations in sensory level from the learned ideal point in laboratory animals (Pierrel, 1958) but triangles were easier to fit to the limited number of data (e.g., Blough, 1967). This identity of the peaked functions for sensed difference (dissimilarity) and motivated behavior (preference) was recognised before sensory evaluation became established (Shepard 1958, 1965).

Data from discrimination scaling of all types are now routinely fitted to a hyperbola with its center at the norm point used by the assessor during the session (Figure 1). Where the context of the testing has been adequately realistic, the hyperbola collapses into the isosceles triangle formed by the intersection of its tangents (Conner and Booth 1991; Conner and others 1986).

Figure 1 about here

When only two stimulus levels are tested, as in a difference test, the responses need to be on one side of the matching point. If there is a physical level on the other side, it should be close to an exact match. The reason is that responses to stimuli far from the peak on both sides may be fitted best by a very shallow line. This is an explanation of the apparent paradox of large differences in strength of stimulation with no difference in preference (Delwich and O'Mahony 1966). The serious consequence of one stimulus level on each side of the peak is that the estimate of the matching point is a long way up or down from the true value, which is seen when more two or more stimulus levels on each side of ideal are tested.

Measures of Strength of Influence

The natural and engineering sciences often use the slope of a function to measure the strength of influence of an input or the sensitivity of an output. Examples include an instrument's calibration line, a drug's dose-response function, or a machine's performance as the percent of optimum. The social and behavioral sciences favor the lack of error in the fit of the input-output function to the data, measured by the (partial) regression coefficient, r (beta).

These two measures can be combined into a single parameter. This generic measure of causal strength has traditionally been called the just-noticeable difference (JND). The root mean square of error in responses in the regression from stimuli to responses is divided by the slope, i.e. response difference divided by stimulus difference (Torgerson, 1958). The response dimension cancels out, leaving only the stimulus dimension. The calculation yields the difference in stimulus levels which the assessor's response levels discriminate with a success halfway from zero (random responding) to perfect (100% correct).

The tendentious term ‘JND’ should be replaced by the phrase ‘half-discriminated disparity’ (HDD), for several reasons. First, for material stimuli, the levels are measured in logarithmic units (equal ratios), whereas for conceptual stimuli the levels remain in the numeric differences (equal intervals) recognised in the culture (Booth and Freeman 1993). Hence the neutral term ‘disparity’ is to be preferred to ‘difference’. Second, the disparity is discriminated within the session that yielded the data, regardless of whether the assessor ‘notices’ it, or is ‘able’ to be aware of the two levels of the stimulus in some other circumstances. Third, ‘just’ distinguishing or not has been taken to imply a discontinuity (threshold) between awareness and unawareness of a disparity, and even a paradoxical awareness of unawareness. To the contrary, the HDD is a point on the smooth continuum of degrees of overlap between the distributions of responses to the two stimuli. It is simply halfway between complete overlap (the two levels being the same) and no overlap (the two levels being infinitely far apart). That is, the unit disparity between levels is where the upper quartile of responses to the weaker stimulus is superimposed on the lower quartile for the stronger stimulus (Conner and others 1988).

A straight line, $y = mx + c$, has two parameters, the slope (m) and the intercept with the axis (c). It may also be specified by two points, (x_1,y_1) and (x_2,y_2). The parameters specifying a particular discrimination hyperbola are its half-discriminated disparity (corresponding to m) and its norm (the personal ideal or the brand matching point, as c).

The observed value of a HDD is the individual’s discriminative sensitivity in the context of testing while using a particular response. If the anchor is the personally most preferred version, then the HDD can be called the ‘just tolerated disparity’ (cp. Conner and others 1988). In either use, the HDD is the fundamental objective measurement of the impact of a sensory (or conceptual) factor on the individual’s response in the context tested.

Tolerated sensory distance is a more subjective measure of personal importance of the tested difference, although still much more operational than mere ratings of “importance” or a similar term. Instead of the two tested levels of a tetrad (or triad) in the sensory distance, the tolerated distance uses the levels at the ‘just wrong’ anchor (folded) or anchors (unfolded). In the case of the present data, these limits of tolerance were insufficient salt (“just too little”) and excessive salt (“just too much”) in the piece of white bread eaten by itself or the spoonful of tomato soup (cp. Prescott and others 2005). These levels are specified by the regression line’s intersections with low and high “never choose” or each unfolded anchor point such as “just too little” and “just too much” (with room for responses at worse extremes).

TAINT OR TREAT

Taints are detected as a difference from the usual product. Yet it may not be correct to assume that the unusual component is aversive to all consumers. The foreign component may increase the attractiveness of the product, at least to some users. A long recognised case is the

boar taint in pork products. At a time when fresh pork from male pigs was available, a minority felt it had a desirably stronger flavor on cooking. For such people, the unusual flavor is a treat, not the taint it is for many. Another example comes from the time when caffeine was extracted from coffee beans using an organic solvent. Those who were sensitive to a sweetish aroma from residues of the solvent felt an extra richness in the flavor, which some found attractive and others repulsive.

Hence averaging responses across the panel misrepresents the market to producers and risks disservice to users by reducing them to the lowest common denominator of dictatorship by the majority. Aggregation of difference test results needs to preserve the direction as well as the size of the sensory effect on preference or familiarity. Rating each sample's intensity relative to ideal and analysing each individual's data provide automatic protection against misinterpretation of the valence of an unusual component. In a tetrad, the two tainted samples will be rated further from ideal than the two usual samples. The duplicate treat will be rated closer to ideal.

A constitutive component of a product can become as aversive as a foreign component if it departs sufficiently from norm. Insufficient flavor is widely recognised as off-putting. Sensory methods have been less effective at protecting product developers against aversive excess of flavorings. This is especially so for sweeteners because they evoke reflex movements of suckling (Booth and others 2010). Arguably the greatest marketing mistake of all time can be attributed to a treat being converted to a taint. New Coke had an aroma flavoring that differed from the longstanding cola. That therefore could have been regarded as a foreign component. Probably the more serious factor in the failure of the new product was excessive sweetness, arising from optimisation tests flawed by the innate reflex, plus range bias (Booth and Shepherd 1988). The monadic test self-administered by a longstanding user of the brand on tasting the new variant was therefore liable to evoke the response that the level of sweetness that was a key part of the delight in cola had been changed to a repulsive over-strength.

In short, difference tests that measure and aggregate individuals' discriminative effects on preference can in principle serve equally well for detection of taints and for optimisation of normal constituents and processing factors (Booth 2014).

The Cardboard 'Taint' in Bread

If sodium ions are eluted from a food at a lower concentration than they are in saliva, this is liable to generate the 'water taste' that is strongest in neutral distilled water (Bartoshuk 1968). Carbohydrate foods that are not appreciably sweetened, such as regular breads in the UK, taste like 'cardboard' at such low concentrations of sodium salts (usually chloride, sometimes bicarbonate as well). Hence a study of the impact of salt levels in bread can serve as a model for taint detection and the effects of increasingly foreign sensory levels on preference.

The earliest report of discriminative acuity of both sensory strengths (how salty) and degrees of preference (closeness to the ideal point) included data on levels of salt in white bread baked for the British market (Booth and others 1983). Most bread in the UK has no added sugar or dried fruit, and is sold in plastic packs of pre-sliced loaves. Reanalysis of those data illustrates how discriminative difference tests within the tolerable range of a sensed constituent can be used to optimise a product. The Figures in this paper present data from the closest pairs of the concentrations of salt in bread that were investigated in that academic-industrial collaboration.

Theoretically there is a continuum from taint or too little, through both sides of the preference peak, and onward to levels of a normal constituent that are too much (Booth 1987). Figure 2 displays this sequence in data that all came from the same tetrad of samples. Only the assessors varied. Their diversity in personal ideal points generated the succession from a positive slope, through a peak, to a negative slope. These results also illustrate how a single tetrad can provide a distribution of discrimination functions which is balanced across the panel between low and high ranges. The two levels in a tetrad for optimisation should be chosen to be within the region of the mode of individuals' ideal points or familiar levels.

Figure 2 about here

Two examples from each of four tetrads of salt in bread are given in Figure 3. The lowest tested two levels of salt were 0.54 and 0.89 g per 100 g of bread (%). The three tetrads including 0.54 g % serve as models of taint testing. (Salivary concentration of sodium ions is around the equivalent of 0.5 g of NaCl per 100 ml; hence those bread samples may on the margin of having a cardboard 'taint'.) The severity of the taint is modeled by size of the contrast of 0.54 with 0.89, 1.5 or 2.4 g % (Figure 3).

Figure 3 about here

As stated above, the HDD is derived from both the slope of the regression line and also the deviations of data points from the line. Hence there is no direct relation between an HDD for a session and either just the slope value (reading units off each of the two axes) or the spread alone of data points around the line (in units of the y axis).

In Figure 3, the slope of the lower graph from the 0.54 vs 0.89 tetrad is four times steeper than the slope in the upper graph, and yet the HDDs are very similar. That is a compensatory effect of relatively small deviates at the lower level of salt in the upper panel.

Conversely, the upper graphs from 0.89 vs 1.5 and 0.54 vs 2.5 have similar slopes, but the HDD is much worse (higher) in the assessor of 0.54 vs 2.5 (Figure 3). That is because of a pair of huge deviates in the latter case. Anecdotally, 2.5 g % is very salty and puts the upper graph's assessor in two minds about liking such samples, while the assessor whose data are in the lower graph finds both 2.5 and 0.54 g % very far from ideal as well as difficult to rate consistently.

This first experiment on sensory discrimination by preference (Booth and others 1983) used characterised distances from the most preferred version, i.e. “saltiness relative to ideal” (e.g., Figures 1, 2 and 3). However, use of an explicit sensory concept is not necessary. Overall preference can be rated from “always choose” to “never choose” and beyond (Booth 2014). Overall match to the remembered familiar version can be rated from “exactly the same” to “unrecognisably different”, for example. Any sensory characterisation can be done on a later run through the samples.

MEASURING THE SIZE OF A DIFFERENCE

Considerable interest in tetrads has been generated by reports that fewer panelists are needed to reach a $p < 0.05$ if two samples of each stimulus are presented instead of two samples of one stimulus and only one of the other (as in a triangle test). However, it has never been clear why a 5% probability of there being no difference should be the criterion for a business decision or for any scientific interpretation. The decision how many panelists are needed is merely an expression of opinion on the importance in that situation of getting conventionally reliable evidence for a difference. The real issue is the objective size of the difference, in terms of its impact on observable action, discussion or thought.

Furthermore, four samples are necessarily capable of providing more information than three samples. More specifically, greater precision in the estimate of error is provided by duplicates of each stimulus than a duplicate of only one of the two stimuli. Least squares regression can be calculated from monadic responses to a difference triad (cf. Figures 5-8 below) but the variance in responses to the singleton is merely assumed to be the same as that to the duplicated stimulus.

The greater statistical power of tetrads (over triads) of course depends also on the nature of the task performed by each panellist. The task depends on two conditions which it is misleading to conflate into the single idea of ‘the question asked’. What the assessor attempts is primarily determined by how the samples are presented. The strategy may also be influenced by the wording of instructions or a response form but this cannot be assumed to occur (e.g. Richardson-Harman and Booth 2006).

There is a long tradition of presenting two samples at the same time and asking for one of two responses that compares the pair in some way (e.g. two alternatives, forced choice: 2AFC). It makes no difference to the discriminative acuity (halfway between random and perfect) or the implicit standard of comparison (either sort of matching point) whether the choice between the two responses concerns sensory intensity or personal preference (McBride and Booth 1987). However the analysis must allow for the contrast between the peak in preferred intensities and the monotonicity of described intensities (MacRae and Geelhoed 1992): a difference in preference can arise from either a decrease or an increase in intensity.

The most economical measure of sensory impact is the effect of two or more levels of stimulation on a conceptualised quantitative response, whether strength of the stimulus or disposition to accept it. The validity of that measure depends on the experimenter's success in simulating the usual context of that stimulation (the "ecological validity" of Brunswick 1955, 1956). Hence neither practical nor fundamental sensory research should isolate the stimulus of interest in an artificial medium or unfamiliar ambience, place and time. Rather, variants of a familiar material should be tested at a time and place as close as has been shown to be relevant to the context of a prevalent use. Adherence to such conditions also helps to avoid common problems such as inadequate amount of each sample (instead of a normal mouthful or used amount), excessive number of samples in a session (satiety), and undue diversity of samples (cheese versus chalk).

Hence the data used in this paper to illustrate discriminative difference testing come from the first sensory experiments to use mouthfuls of familiar foods, eaten close to a mealtime at a table near a kitchen using regular utensils and limiting the total amount consumed to within the usual portion size (Booth and others 1983). Sensory levels were selected in subsequent experiments to avoid biases on intensities and preferences that arise from levels that range high or low (Conner and others 1986; Riskey and others 1979). Responses positioned each test stimulus on a straight line specified by the main anchor on the optimum level (just right) and effectively the minor anchor of intolerably far from optimum. In fact, these early experiments unfolded the limit on personal tolerance into too little and too much. Nevertheless, assessors forced those two extreme levels into the same distance from the optimum level: there was no reliable difference in panel means between the regression slopes below and above optimum (Conner and Booth, 1992). Subsequent work used folded responses from just right to just wrong (see Booth 2014).

Sensory Distance

The data from each assessor provide an estimate of the sensory distance between the tetrad's two levels. This perceived disparity between two levels of salt (or whatever is the sensed factor) is measured in units of discrimination (HDDs). For example, as the ratio of the level contrasted to 0.54 g % increases, so does the sensory distance between the two levels tested.

This distance depends on both the size of the physical ratio of salt concentrations and also the individual's discriminative acuity during the session (the HDD). Hence, the general theoretical relationship is subject to variations among individuals in the performance of differential acuity between the two levels actually presented. High acuity (a low HDD value) will increase the number of HDDs at any ratio of physical levels. An unusually large HDD (poor discriminative acuity) will make the two levels seem closely similar. Plotting panel-median discriminative distances against tetrad (and triad) ratios confirms that the individual performances in the aggregate improve systematically as the physical disparity increases (Figure 4).

Figure 4 about here

This basic sensory distance is the most objective measure available of the importance of the difference to the assessor. In the case of tetrad of a tainted sample and the usual untainted version, the number of HDDs between the samples is the functional size of the taint.

AGGREGATION ACROSS PANELISTS

Measurements of each panelist's performance can readily be aggregated across the panel, to provide the generalisation required about a taint or about the optimisation of a constituent for the market or a segment of it. This paper is based on tetrads for all the pairs of closest salt levels sampled in duplicate in the raw data summarised by Booth and others (1983). In addition, triads were derived from these tetrads, together with individuals who had only one sample at a level adjacent to a level tested twice.

Distribution of Discriminative Differences

The frequency polygons for half-discriminated disparities are aggregated across the panels tested on each tetrad or triad in Figure 5. There are physiological limits on differential acuity and so HDDs tend to a minimum. Any interfering factors reduce that acuity, giving larger HDDs. Great interference is less likely and so the distribution follows a (reverse) J curve. With sufficient data, these distributions are amenable to survival analysis, with the possibility of identifying distinct sources of interference with fine discrimination.

Figure 5 about here

With the limited number of salt levels in the bread, the triads appeared to be more susceptible than the tetrads to interference with discriminative performance: the J curve fell off more gradually (top panels, Figure 5). The measure may have more susceptible to lack of sampling of one of the two levels. Hence tetrads would be better than triads for taint measurement, and also for optimisation when rather few variants of a sensory or conceptual factor are available.

On the other hand, in both bread and soup, triads may have been better than tetrads at pushing assessors towards the limit of performance: the mode was only at the second bin ($0.10 < HDD < 0.19$) for tetrads (Figure 5). This may be a consequence of another sort of sampling effect: data from tetrads have a greater chance of including a pair of highly disparate responses to a duplicated level.

The whole set of samples tested in each individual ran from below to above the personal optimum, in order to minimise range bias (Booth and others 1983). Hence, it was possible to compare tetrads and triads with both levels below the optimum and with both levels above the optimum. Furthermore the triads selected from a tetrad could have the odd one out at the extreme or closer to the other pair in the trio. There were enough triads of the soup to split the data these four ways (bottom panels, Figure 5). It appeared that pressure to the limit of discrimination arose from the level sampled once only being at an extreme ("very low" or

"very high" in Figure 5). Assessors may be less attentive when a sample is closer to the ideal point.

Distribution of Ideal Points with Tolerance Ranges

When the assessor is using the most preferred level as the norm of comparison for each sample, the HDD is the range of tolerance of deviations from that ideal point. These two parameters can be combined for each assessor in a norm range [one HDD on either side of a norm point (NP), such as the ideal level]. Within this range, deviations from ideal are less than half discriminated.

An ideal range (the ideal point combined with the HDD) indicates how important the precise value is to the panellist, i.e., how tolerant s/he is of deviations from it. Those ideal points that are most precise and potentially influential in a competitive market are represented by sharper elevations in the profile of counts (Figures 6 and 7).

Figures 6 and 7 about here

OPTIMISATION

Such aggregations of panelists' individual data can be used to optimise the sensory or conceptual factor in preference -- the salt content of the bread or of the soup in this illustrative case. Data from tetrads (or triads) must be used with care for this purpose, because the two levels need to be either below or above the matching point (to the personal ideal or to the target product), although not excluding one level close to match on either side (cp. the tetrad below ideal in Figure 1 and three assessors in Figure 2). Effectively identical submodes below and above ideal were observed (Figure 6) but below ideal also had a lower submode and above ideal a higher one. Such range biases can be compensated by combining data from equal numbers of assessors observed to have been tested below and above their ideal points. This combination of selected difference data then gives a similar profile to that composed of all the data on bread collected by Booth and others (1983) (bottom panel of Figure 6).

When the soup triads were split among four pairs of salt levels, a single mode of ideal discrimination ranges appeared at each level, and the four modes were effectively identical (Figure 8). This finding supports the view that the width of the range in a two-level discrimination test is not critical, unlike its position relative to the personal ideal or target product.

Market Response Profiles

The realism and power of discriminative difference testing are illustrated by the capacity of data from a panel to show directly the response of the market which is represented by the procedure for recruiting panelists.

Members of a panel should be selected in accord with the purpose of the investigation, not from convenience or tradition. If the aim is to match an existing product, or more basically to understand how a particular food works, then the panel can be anyone who is familiar with the product. If the aim, rather, is to estimate the response of an existing market, then the panel has to be a representative sample of users. If the product has more than one distinct use, then a panel and test design is needed for each use. In either case, as in any science, the testing conditions should mimic the conditions of use as closely as feasible.

Positioning of brands or varieties of a brand to discriminative segments can be considered if there is more than one major mode in the distribution of ideal points, or if the distribution is wide enough for positions in its two wings together to include more people than one central position does. The range of one HDD on either side of the ideal indicates when the level in the marketed product could be distinguished from ideal by the user in the situation simulated by the test. However such differences might well be tolerated. Indeed, a user's ideal point is likely to adapt to the new level as it becomes familiar. To accommodate this possibility, tolerated discrimination distances can be substituted for HDDs. That generates distributions with much more overlap between individuals. A marketed level decided from such a distribution will be within the range of initial acceptance of a larger proportion of the market.

Multiple Sensory and/or Conceptual Factors

Designs that keep close enough to each assessor's multiple-factor ideal point or target match extend to any number of factors tested with at least two levels minimally correlated with variation in other factors. This is because each factor forms its own discrimination hyperbola. The ideal or the tolerance range can therefore be extracted for each factor for application to the whole product or brand (for examples, see Booth 2014).

The crucial distinction between this approach and established practise is that the performance characteristics of each panelist are calculated from the raw data in accord with scientific theory before any aggregation across the panel is attempted. This contrasts with the application of distribution-free or normal statistical models to the whole panel's raw data first (e.g., Næs and others 2014), with or without individualisation (e.g., Jaeger and others 2000).

CONCLUSION

This paper presents the simplest possible example of a scientific approach to applied sensory studies. The theory and practise are long established but remain innovative.

A fundamental divide between sensory analysis and analysis of consumer preference continues to be claimed in principle and implemented in practise. That thesis was shown to be untrue three decades ago (Booth and others 1983; Conner and others 1986; McBride and Booth 1986). The present paper carries that refutation through to the fundamental

mathematics and the procedures needed to formulate an agreed standard of applied sensory research on preferences and perceptions. Sensory analysis relevant to the supplier of a product needs to be done with users of the existing (sub)brands, on those products and any new propositions needed to test the factor(s) thought to be of importance to a business in the supply chain. The impact of a sensory difference on choices in the market should be estimated from data on representative individuals, which have been collected before the investigator draws attention to sensory or marketed concepts. All such work should be driven by client-relevant hypotheses, not by statistical models or diagrams that have previously interested business people.

Panels using specified concepts to compare pairs of samples have more statistical power with tetrads of samples than with even two replications of sample dyads (Garcia and others 2013; see also Ishii and Mahony 2014). However, such observations are not relevant to the use of monadic judgments on sample tetrads to measure each panelist's perceived strength of a taint or physical value for ideal strength or match to the target. Indeed, it would entirely miss the point of this paper to ask about the statistical power of sample tetrads in the discriminative measurement of a difference. Existing sensory standards use procedures of data analysis that are incapable of measuring the size of a perceived difference or its impact on a product user's choices, concepts or sensations. That requires a move from probabilistic evaluation of data to scientific measurement of what is actually happening.

REFERENCES

BARTOSHUK, L.M. 1968. Water taste in man. Perc. Psychophys. *3*, 69-72.

BLOUGH, D.S. 1967. Stimulus generalization as signal detection. Science *158*, 940-941.

BOOTH, D. 1987. Individualised objective measurement of sensory and image factors in product acceptance. Chem. Ind. (Lond.) (1987 Issue 13), 441-446.

BOOTH, D.A. 1988a. Estimating JNDs from ratings. [Abstract, AChemS-10] Chem. Sens. *13*, 675-676.

BOOTH, D.A. 1988b. Practical measurement of the strengths of actual influences on what consumers do: scientific brand design. J. Market Res. Soc. (U.K.) *30*, 127-146.

BOOTH, D.A. 1994. Psychology of nutrition, Psychology Press, Hove UK.

BOOTH, D.A. 2009. Lines, dashed lines and "scale" ex-tricks. Objective measurements of appetite *versus* subjective tests of intake. Appetite *53*, 434-437.

BOOTH, D.A. 2014. Measuring sensory and marketing influences on consumers' choices among food and beverage product brands. Trends Food Sci.Technol. *35*(3), 129-137.

BOOTH, D.A., and CONNER, M.T. 2009. Letter to the Editor [about salt in bread]. J. Food Sci. *74*(3), vii-viii.

BOOTH, D.A., and FREEMAN, R.P.J. 1993. Discriminative feature integration by individuals. Acta Psychol. *84*, 1-16.

BOOTH, D.A., and FREEMAN, R.P.J. 2014. Mind-reading versus neuromarketing: how does a product make an impact on the consumer? J. Cons. Marktg *31*(3), 177-189.

BOOTH, D.A., HIGGS, S., SCHNEIDER, J., and KLINKENBERG, I. 2010. Learned liking versus inborn delight. Can sweetness give sensual pleasure or is it just motivating? Psychol. Sci. *21*, 1656-1663.

BOOTH, D.A., O'LEARY, G., LI, L., and HIGGS, S. 2011a. Aversive viscerally referred states and thirst accompanying the sating of hunger motivation by rapid digestion of glucosaccharides. Physiol. Behav. *102*, 373-381.

BOOTH, D.A., MOBINI, S., EARL, T., and WAINWRIGHT, C.J. 2003. Consumer-specified instrumental quality of short-dough cookie texture using penetrometry and break force. J. Food Sci.: Sens. Nutr. Qual. *68*(1), 382-387.

BOOTH, D.A., SHARP, O., and CONNER, M.T. 2011b. Discrimination without description of differences. Implicit or fully subconscious? http://epapers.bham.ac.uk

BOOTH, D.A., and SHEPHERD, R. (1988). Sensory influences on food acceptance - the neglected approach to nutrition promotion. BNF Nutr. Bull. *13*(1), 39-54.

BOOTH, D.A., THOMPSON, A.L., and SHAHEDIAN, B. 1983. A robust, brief measure of an individual's most preferred level of salt in an ordinary foodstuff. Appetite *4*, 301-312.

BOWMAN, S.J., BOOTH, D.A., PLATTS, R.P., and UK Sjögren's Interest Group. 2004. Measurement of fatigue and discomfort in primary Sjögren's syndrome using a new questionnaire tool. Rheumatology *43*, 758-764.

BRUNSWIK, E. 1955. The relation of the person to his environment. Acta Psychol. *11*, 108-112.

BRUNSWIK, E. 1956. Perception and the Representative Design of Psychological Experiments. 2nd ed. University of California Press, Berkeley.

CARROLL, J.D. 1972. Individual differences and multidimensional scaling. In *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences,* Volume 1 (R.N. Shepard, A.K. Romney and S.B. Nerlove, eds.), Seminar Press, New York.
CARROLL, J.D., and CHANG, J.J. 1970. Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrika *35*, 285-319.
CHECHLACZ, M., ROTSHTEIN, P., KLAMER, S., PREISSL, H., PORUBSKA, K., HIGGS, S., BOOTH, D., ABELE, H., BIRBAUMER, N., and NOUWEN, A. 2009. Diabetes dietary management alters responses to food pictures in brain regions associated with motivation and emotion: an fMRI study. Diabetologia *52*, 524-533.
CONNER, M.T., and BOOTH, D.A. 1991. Characterisation and measurement of influences on food acceptability by analysis of choice differences: theory and practice. Food Qual. Pref. *2*, 75-85.
CONNER, M.T., and BOOTH, D.A. 1992. Combining measurement of food taste and consumer preference in the individual: reliability, precision and stability data. J. Food Qual. *15*, 1-17.
CONNER, M.T., BOOTH, D.A., CLIFTON, V.J., and GRIFFITHS, R.P. 1988. Individualized optimization of the salt content of white bread for acceptability. J. Food Sci. *53*, 549-554.
CONNER, M.T., BOOTH, D.A., and HADDON, A.V. 1987. *Individualised analysis using free-choice descriptor identifies sensory preference segments.* Unpublished manuscript of a talk to the Sensory Panel of the Food Group, Society for Chemical Industry, London UK.
CONNER, M.T., HADDON, A.V., and BOOTH, D.A. 1986. Very rapid, precise measurement of effects of constituent variation on product acceptability: consumer sweetness preferences in a lime drink. Lebens. Wiss. Technol. *19*, 486-490.
COOMBS, C.H. (1964). A Theory of Data, Wiley, New York.
DELWICH, J., and O'MAHONY, M. (1996). Flavour discrimination: an extension of Thurstonian 'paradoxes' to the tetrad method. Food Qual. Pref. 7(1), 1-5.
FECHNER, G.W. 1860. Elemente der Psychophysik. Volumes 1 & 2. Breitkopf & Hartel, Leipzig. Translated into English by H.E. Adler (1966), *Elements of Psychophysics* (D.H. Howes and E.G. Boring. eds.), Holt, New York.
GARCIA, K., ENNIS, J.M., and PRINYAWIWATKUL, W. 2013. Reconsidering the specified tetrad test. J. Sens. Stud. *28*, 445-449.
GARITTA, L.V., SERRAT, C., HOUGH, G.E., and CURIA, A.V. 2006. Determination of optimum concentrations of a food ingredient using survival analysis statistics. J. Food Sci. *71*, S526-S532.
GOODCHILD, C.E., TREHARNE, G.J., BOOTH, D.A., KITAS, G.D., and BOWMAN, S.J. (2008). Measuring fatigue among women with Sjögren's syndrome or rheumatoid arthritis: a comparison of the Profile of Fatigue (ProF) and the Multidimensional Fatigue Inventory (MFI). Musculoskeletal Care *6*, 31-48.
ISHII, R., and MAHONY, M. 2014. Triangle and tetrad protocols: small sensory differences, resampling and consumer relevance. Food Qual. Pref. *31*, 49-55.
JAEGER, S.R., WAKELING, I.N., and MACFIE, H.J.H. 2000. Behavioural extensions to preference mapping: the role of synthesis. Food Qual. Pref. *11*, 349-359.

JONES, L.V., PERYAM, D.R., and THURSTONE, L.L. 1955. Development of a scale for measuring soldiers' food preferences. Food Res. *20*, 512-520.
LÓPEZ OSORNIO, M.M., and HOUGH, G. 2010. Comparing 3-point versus 9-point just-about-right scales for determining the optimum concentration of sweetness in a beverage. J. Sens. Stud. *25*, 1-17.
MACRAE, A.W. 1995. Confidence intervals for the triangle test can give reassurance that products are similar. Food Qual. Pref. 6(2), 61-67.
MACRAE, A.W., and GEELHOED, E.N. 1992. Preference can be more powerful than detection of oddity as a test of discriminability. Percept. Psychophys. *51*, 179-181.
MCBRIDE, R.L. 1983. A JND-scale – category-scale convergence in taste. Percept. Psychophys. *34*, 77-83.
MCBRIDE, R.L., and BOOTH, D.A. 1986. Using classical psychophysics to determine ideal flavour intensity. J. Food Technol. *21*, 775-780.
MORGAN, M.J., WATAMANIUK, S.N., and MCKEE, S.P. 2000. The use of an implicit standard for measuring discrimination thresholds. Vision Res. *40*, 2341-2349.
NACHMIAS, J. 2006. The role of virtual standards in visual discrimination. Vision Res. *46*, 2456-2464.
NÆS, T., TOMIC, O., GREIFF, K., and THYHOLT, K. 2014. A comparison of methods for analyzing multivariate sensory data in designed experiments. A case study of salt reduction in liver paste. Food Qual. Pref. *33*, 64-73.
PIERREL, R. 1958. A generalisation gradient for auditory intensity in the rat. J. Exp. Anal. Behav. *1*, 303-313.
PRESCOTT, J., NORRIS, L., KUNST, M., and KIM, S. 2005. Estimating a "consumer rejection threshold" for cork taint in white wine. Food Qual. Pref. *16*, 345-349.
RISKEY, D.R., PARDUCCI, A., and BEAUCHAMP, G.K. 1979. Effects of context on judgments of sweetness and pleasantness. Perc. Psychophys. *26*, 171-176.
RICHARDSON-HARMAN, N.J., and BOOTH, D.A. 2006. Do you like the sight or the feel of milk in coffee? Ecology and effortful attention in differential acuity and preference for sensed effects of milk substitute in vended coffee. Appetite *46*, 130-136.
ROSS, H.E., and MURRAY, D.J. (eds.) 1996. E. H. Weber on the Tactile Senses, Psychology Press, Hove, U.K.
SHEPARD, R.N. 1958. Stimulus and response generalization – tests of a model relating generalization to distance in psychological space. J. Exp. Psychol. *55*, 509-523.
SHEPARD, R.N. 1965. Approximation to uniform gradient of generalization by monotone transformation. In D.I. Mosofsky (Ed.), *Stimulus generalization*, pp. 94-110, Stanford University Press, Stanford CA.
STEVENS, S.S. 1961. To honor Fechner and repeal his law. A power function, not a log function, describes operating characteristic of a sensory system. Science *133*, 80-86.
STEWART, N., BROWN, G.D.A., and CHATER, N. 2005. Absolute identification by relative judgment. Psychol. Rev. *112*, 881-911
THOMSON, D.M.H., and CROCKER, C. 2014. Development and evaluation of measurement tools for conceptual profiling of unbranded products. Food Qual. Pref. *33*, 1-13.
THURSTONE, L.L. 1927. A law of comparative judgment. Psychol. Rev. *34*, 273-286.
TORGERSON, W.S. 1958. Theory and methods of scaling, Wiley, New York NY.

WEBER, E.H. (1834/1996). *De pulsu, resorptione, auditu et tactu* [On the pulse, breathing, hearing and touch]. Translation by H.E Ross & D.J. Murray (Eds.), *E.H. Weber on the tactile senses*, pp. 23-136, Psychology Press, Hove UK.
WILLIAMS, A.A., and ARNOLD, G.M. 1985. A comparison of the aromas of six coffees characterised by conventional profiling, free-choice profiling and similarity scaling methods. J. Sci. Food Agric. *36*, 204-214.

Titles of Figures, with subtitles and footnotes

FIG. 1. THE PSYCHOPHYSICAL HYPERBOLA FOR EACH ASSESSOR

Each graph shows the hyperbolic best fit to the data from the assessor having the panel's median HDD value for the tetrads or triads below or above the personally most preferred salt levels in white bread without spread.

Note. The continuous line (in a light color) is the least-squares best fitting hyperbola, with a peak that is rounded to some extent by contextual defects (negligible in these instances). The broken line (in a darker color) is a tangent to the hyperbola, i.e. the back extrapolation of one of its asymptotes to the intersection with the other tangent at the hyperbola's centre (the apex of the isosceles triangle).

log NaCl (S1): the first and only Stimulus, sodium chloride, at concentrations of grams per 100 g of bread, as a logarithm to the base 10.

salty rti (R1): the acceptance Response to the characterised attribute of NaCl, "salty". The plotted score would be zero if a response were placed on the anchor category of just as salty as liked. Plotted score of -50: so far from ideal as to be intolerable, i.e., either just too little or just too much for the assessor to choose.

Codes at top of graph: assessor's numerical name and tetrad or triad category, description of sampled material, replication number, investigator's initials and source document (the student's report on a research project). (Graphics output from a calculator of cognitive processes, including those in the appreciation of a consumer product, Co-Pro2.29)

FIG. 2. RESPONSES BY SIX ASSESSORS TO A SINGLE TETRAD (1.5 *versus* 2.5 g / 100 g; in logs = 0.18 *vs* 0.4)

FIG. 3. TETRADS OF SALT IN BREAD, WITH EACH SAMPLE RATED FOR DISTANCE FROM PERSONALLY MOST PREFERRED SALTINESS

Notes. Raw data from Anne L. Thompson's BSc project report, summarised in Booth, Thompson and Shahedian 1983; Booth 2014).

Each graph shows the four data points (x) from one assessor (A#) for one of four selected tetrads, three of which included two samples of 0.54 g NaCl per 100 g bread (%) and two samples of 0.89%, 1.5% or 2.5%.

x axis: concentration of Stimulus (S1) in $\log_{10}$ g of sodium chloride in 100 g of bread loaf (a production variable, allowing for water lost during baking, not an instrumental value for crumb).

y axis: score of Response (R1), line position for how salty relative to ideal (rti).

Midpoint "saltiness just right" = 0; endpoints "not nearly salty enough" or (folded) "much too salty" both plotted at -50.

Continuous line: hyperbolic regression forced through a peak score of zero, rated as the ideal salt level.

Broken line: tangent to the fitted hyperbola.

Graphics output from runs of the calculator program Co-Pro 2.29 (Booth, Sharpe, Freeman & Conner 2010/1).

Distance: number of HDDs between the two tested levels of salt.

HDD ratio: higher over lower g / 100 g at 50% discrimination (one plus the Weber fraction).

Ideal point: salt level interpolated to a "just right" response.

FIG. 4. RELATIONSHIP OF DISCRIMINATION DISTANCE (NUMBER OF HALF-DISCRIMINATED DISPARITIES) TO THE RATIO OF STIMULUS LEVELS WITHIN A TETRAD, FOR SALT IN WHITE BREAD OR TOMATO SOUP

Note. Many more levels of salt were available in the laboratory-prepared samples of soup than in the manufacturer-provided samples of bread (Booth, Thompson and Shahedian 1983). Hence far fewer tetrads could be extracted from the original data collected on bread.

FIG. 5. INCIDENCES OF VALUES OF THE HALF-DISCRIMINATED DISPARITY (HDD) FROM TETRADS AND TRIADS OF SALT LEVELS IN WHITE BREAD AND TOMATO SOUP

Notes. Regressions with $r^2 < 0.4$ were excluded. There was a large enough total of soup triads to split them by the unique sample (odd one out) being lower or higher than the duplicated sample, with much lower or higher odd values being further separated out. The mode of half-discriminated fractions remained below 0.1 in all subsets of triads, but the less extreme singletons seemed to give a higher incidence of estimates of a moderately less acute HDD (0.1 to 0.3).

FIG. 6. COUNTS OF NORM RANGES FOR SALT IN BREAD FROM TETRADS AND TRIADS BELOW AND ABOVE IDEAL, PLUS ALL THE DATA FOR EACH ASSESSOR

FIG. 7. IDEAL DISCRIMINATION RANGE COUNTS FROM ALL TETRADS (UPPER PLOT) AND TRIADS (LOWER PLOT) OF TOMATO SOUP

FIG. 8. IDEAL DISCRIMINATION RANGE COUNTS FOR TRIADS OF SALT IN SOUP AT DIFFERENT PAIRS OF LEVELS

Note. The scales of salt level (x axes) are approximately equated, aligned vertically and matched horizontally.