The basics of quantitative judgment. How to rate the strength of appetite for food and its sating.

DAVID A. BOOTH

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

Abstract

The current strength of a person's appetite for food can be observed in any graded expression of the disposition to take a mouthful of food. For this quantitative judgment to measure an influence on hunger/satiety, however, the source of that influence has to be varied independently of other influences at the moment the rating is made.

Key words: quantitative judgments; ratings; scores; magnitude estimates; scales

Corresponding author's *Email address:* D.A.Booth@Bham.ac.UK (David Booth)

A person's present appetite for food and drink

Performance, not experience

A good rating of appetite is easy to elicit. The investigator asks for an indication of the current tendency to eat, in a way that can be scored from zero for no appetite, to nine (or ten, or seven) for "would always eat." The question can be narrowed to a particular food, whether named or presented for viewing or tasting. Alternatively the question can be broadened to the consumption of any food or drink.

Quantitative judgments of appetite have been made to seem very difficult and dubious by a delusion from before the rise of science in the eighteenth century. That notion still pervades fields outside psychology and lingers in some parts of psychology itself, even though mainstream research was disabused of the worst forms of the error over half a century ago. The mistake is thinking that there is any way of measuring the intensity of a sensation by obtaining a number in answer to a single question or indeed that, even if a score could be attached to a private experience, it would be of any general scientific interest or practical use. Appetite is not at all subjective.

What matters to the understanding of eating and drinking, to effective supply of foods and drinks, and to improving their uses for health, wealth and happiness, is the measurement of one or preferably two or more simultaneous influences on the momentary tendency to eat, as they work consciously and unconsciously though the mind of an individual. A rating of appetite is a person's mental performance on the context in which it is elicited.

That is, appetite is an objective public achievement, not a subjective private experience. The challenge is adequate design of the situation in which a rating is collected. Working out what a person would do in those circumstances is not the real problem.

Wanting food

The disposition or desire to eat (or not) is an observable state of mind that varies from moment to moment, rapidly during eating and usually more slowly between such occasions. The intensity of this ingestive appetite varies from raging hunger to a deep revulsion against food -- most commonly, from some attraction to food and drink after a period without their consumption to degrees of satiety during or between meals, induced by the recent eating or drinking. The momentary motivation can also be selective among the foods or drinks most wanted or unwanted, i.e. material-specific appetite or satiety.

Two major scientific implications of the fact that appetite is a transient mental state have been almost universally neglected in research into food choice and intake.

One implication is that any concept or assessment of appetite that is not of the individual's state at the instant has to be theoretically and operationally defined in terms of sets of such states. The palatability of a food and the satiating power of a nutrient, for example, are scientifically meaningless concepts without specification of the combination of momentary states on which either is based – levels of appetite under what conditions when and where (Booth, 1990, 2009).

The other neglected implication is that the measurement of a state of appetite has to include both a consumption-relevant self-expression by the potential eater and also observed sources of influence on that expression: a mental state is nothing more or less than the individual turning some information from the external and internal

environments into another pattern of information that is or could be imposed on those surroundings (Booth, 2008; Booth & Freeman, 1993).

Such neglect of basic science is of a piece with thinking that the rating of appetite is filling in a fixed "questionnaire." Scores from answers to a set of questions add nothing to knowledge unless they compare states that are equivalent except in, at the very least, one manipulated aspect. In appetite for food, that difference might be in levels of a sensed constituent, in the actions of ingested materials on visceral signals to the brain or in cultural roles as signalled in beliefs about the calories in the foods compared or by a pack label or advertising theme (Booth *et al.*, 1982, 1983; Freeman *et al*., 1993; Kissileff *et al*., 2008).

It is irrelevant what the investigator thinks that the answers to the questions mean. All that matters in a report of the study is the evidence on what each person actually had in mind at the two or more comparable moments. The difference in scores is determined by the whole context of questioning, not just by the wording and display for each rating. Yet no two sets of circumstances are ever the same, even in the same person at different times in exactly the same laboratory session or everyday routine. Hence, abstract arguments about the validity, accuracy or reliability of ratings of a state such as appetite are absurd. Without a particular theoretically or practically formulated purpose, investigation of the reproducibility of a rating is pointless.

Quantitative performance in the answering of a question

Valid measurement of the state of appetite

The basic principles of the use of ratings in scientific measurement are straightforward. Successful ratings achieve quantitative judgments of observable realities. That is, the scores do not give privileged access to the influences on them. Whether the rater has been able to pick out an amount in the world is an issue about which a question and its answer in isolation tell us nothing. If the question and answer make any sense to the rater, they are about something -- in this case, directly or indirectly about eating some food -- but what quantity influences the score is unknown unless the investigator has constructed the situation to provide some evidence.

Unfortunately nearly all users of ratings of appetite since the 1980s have failed to provide even the first step in such evidence, which is psychometric validation of an appetite score by low correlations with scores from answers of other questions. Yet this identifies only the type of effect -- that is, the strength of appetite for food (hunger) or of its sating, rather than appetite for fluid (thirst) and its quenching, or some abnormal suppressant of hunger, such as abdominal discomfort, nausea or acute anxiety.

Experiments with designs that show one or more specific influences on appetite, even when psychometric construct validation is not reported, can provide evidence that different wordings for hunger produce highly correlated scores. This is clear in a table showing all the mean differences between conditions at the point in time when the influences were operative (e.g., Booth *et al*., 1982). The same is clear from presentation of graphs for the different wordings that show identical profiles of scores over time, but this wastes lots of space and exposes poor design.

Furthermore, such tracking of one or more ratings of appetite between meals or within meals cannot in principle tell us anything more than any one wording of rating at the moment(s) of known action of an influence could do. Profiles of ratings

within meals (Yeomans, 2000), like rates of intake (Booth, 2009), merely replicate the general satiating effect of the past-eaten food during its digestion (Booth *et al*., 1982) plus the fading of facilitation of eating by unmeasured combinations of influences such as food-specific habituation, cultural norms for sizes of portions and the onset of generic loss of interest in eating as a result of the test food's actions in the stomach and beyond.

Comparisons of the full variety of question wordings and response layouts in the mid-1970s showed that neither vocabulary nor format affected the direction of effect on rated appetite. The approach that showed the most reliable grouped differences between less and more sated appetite for food was to assess the disposition to eat a familiar item. In the later work by Booth *et al*. (1982), the score that was most sensitive to either somatic or social inhibition of eating was the average pleasantness at the moment of rating of the consumption of a small portion of one of several staple foods. The mean across the named foods of the amount of each item that the rater wanted to eat on its own was slightly less sensitive, maybe because weights of quite different foodstuffs were averaged. When how much would be eaten was judged for only a food that might be eaten at the time of questioning (between meals), this score was exquisitely sensitive to the postingestional effects of fats and slow-release starch when they are generating signals of satiety (Dibsdall *et al*., 1996).

The appetite hypercone

Ratings of appetite are peaked on each of the influences on eating, whereas perceived strengths of those influences are monotonic. This basic mathematical principle of psychology was recognised by Coombs (1964) and incorporated in the non-metric statistical programs (Carroll & Chang, 1970) that are now used for the automated collection and analysis of sensory evaluations to be related to consumers' preferences among foods. Yet in sensory consumer research, even by psychologists, ratings of preference are almost always averaged across individuals in plots against the food samples' contents of sugar or fat. Averaging ratings of intensity makes some sort of sense because each person's data lie on a continuously rising line, as do the amounts of the sensed material. Averaging preferences, however, ignores the peaked relationship between the individual's preference and the amount of sensed material and so blunts the sharpness of the distribution of individuals' most preferred levels into a broadly rounded shape.

When the discrimination mechanism at the basis of all learnt influences on appetite is measured in an otherwise personally ideal context, rated preferences plotted against ratios of a physical measure of the sensed source of influence fall on an isosceles triangle with its peak at the most preferred level (Booth *et al.*, 1983; Conner & Booth, 1988). This implies that the effects on appetite of two influences (sensed or conceptualised) are on the surface of a cone, and three or more influences plot as an unvisualisable 'hypercone' (Booth & Freeman, 1993). This theory applies to all ratings of acceptability, satiety and any other response where the level of an influence can be too strong or too weak.

Hence preference (or aversion) ratings should only be averaged across people (mean, median or mode for the group) if applied to one particular object, name or situation. If variation in a characteristic across the assessed objects is being investigated, then the data must be used to estimate the level of the characteristic that is most liked or disliked by each individual and the frequencies of those maximally preferred or rejected values graphed and analysed.

Scales, scales and scales

The word 'scale' is used for three very different things involved in rating or quantitative judgement. The widest use is for the physical layout on which a rater puts the quantitative response to the question being asked, analogously to the marks on the dial of a meter. Because of the risks of confusion with scientific use of the word, it would be better if this use of the term in ordinary life were not transferred into research.

The fundamental scientific use of the term in the present context is the psychological scale. This is an objective mental performance of transforming different levels of an input into output on a layout for quantitative responding.

A third sense refers to a set of questions to which respondents give graded answers that are highly correlated: this is known in psychometrics as a scale, or a subscale where a whole inventory of questions (each with its own 'scale' for responses) factor-analyses into several multi-item scales (Nunnally & Bernstein, 1994). Successful construction of a psychometric scale identifies the existence of a psychological scale but does nothing to show what it is a scale of. That requires additional forms of validation, the weakest being discriminative validity, where multi-item scale scores are dissociated from each other, and the strongest being quantitative validation on an external criterion, sometimes called psychophysical validation (Booth, 1995). Because the multi-item score is part only of one method for measuring a psychological scale, this use of the term should be regarded as parasitic on the basic scientific meaning.

Symbols of appetite

Influences on appetite cannot be measured merely by the wording of ratings, at least not without psychophysical validation for the exact personal circumstances of those judgments. The words on a dial are misleading if the instrument is not correctly connected to the source of information to which those words to refer. A driver may put a definite interpretation on the reading from a meter in front of him, such as showing the fuel tank to be over three-quarters full, but he is dangerously deluded when that meter in fact performs the role of a temperature gauge and the engine is overheating.

That is, the 'scale' on a thermometer or a map only works because it is part of a measuring instrument, designed to provide readings on the scale (using the term in its proper basic sense) of temperature or distance – in degrees Celsius or kilometres. Readings on a meter, millimetre markings on a ruler and scores from rating on a line or by numbers are meaningless unless the rating procedure is part of a design that measures something, i.e. delivers a position on a real psychological scale – a specific quantitative achievement by the mind of a person operating in an observed context. So it is unacceptable for a graph, table or text in a report to label a rating as "hunger" or "pleasure" unless those words are what rater saw (and therefore should be placed between quotation marks, at least in Method).

Furthermore, in order to claim that the study has measured appetite for food, the investigator must show that the scores come from that hunger motivation or its sating (inhibiting appetite for food by eating) and not from thirst (appetite for water), relief from distress, greed, pleasure or general excitement, for example. A bane of research on eating and drinking for 40 years has been confusion between sensual pleasure from food and the desire to consume it, i.e. appetite (Booth, 1991). This muddle started with the false implication that satiety is a loss of pleasure. It has

recently got worse with unfounded claims to have shown in people the separation demonstrated in rats of the movements involved in swallowing any food from the movements of the mouth that are specific to the taste of sugars (Wyell & Berridge, 2004). These fallacies are based on the twin assumption that "liking" for a food and how "pleasant" it is to consume are purely sensory and involve pleasure (are 'hedonic' ratings). This flies in the face of the language that participants bring to the task of rating. "I'd like that food right now," "It would be pleasant to eat some of it" and "I want to eat that food" all mean exactly the same: the food is appetising to me at present. Even more to the point, it has been shown repeatedly that ratings of "liking" for a food or drink correlate as highly with other ratings of appetite in other wordings as those ratings do with each other. Rating how "pleasant" are some ordinary (quite bland) foods is actually the most sensitive measure of both generic and food-specific satiety within and between meals (Booth *et al*., 1982; Rolls *et al*., 1981).

Thus each question in the context of its posing and the answers as scored constitute an experiment that gives findings that are entirely dependent on its design – the manipulation and monitoring the conditions, the environmental and behavioural references of the question and the linearity of the scored responses relative to the influence assessed. The rater's performance in the task of assessing appetite-related state(s) can only be assessed in terms of the discriminative sensitivity of a particular rating to a putative influence on appetite and to the act of eating (or refusing food) at a particular time and place (Booth & Freeman, 1993).

Symbols of quantity

In order to design an effective rating, it must be appreciated that the task is to express a judgment as to the quantity of some observable, such as taking or declining to take a mouthful of a food or drink, by identifying a position on a straight line representing that quantity (where 'line' does not exclude a series of numbers, even implicit). This spatial task is impossible unless two points on the line are specified in a way that refers to the same observable quantity (not to two different ones). It is impossible to be precise unless the words, pictures or other symbols of these two anchor points can be used by the rater to identify the two same values of the observable on each occasion that the rating is made.

We started comparing wordings and formats to express strength of appetite by rating hunger pangs, sensation of fullness, pleasantness and amount wanted of a staple food, desire to eat and so on as percentages (0-99) in the mid-1970s and shortly after introduced lines of 20 dashes (scored by counting, not in millimetres from a ruler). We have successfully used 6 or 8 boxes or the digits 0-9 between anchors instead of a continuous line for ratings of different facets and domains of pain, discomfort and fatigue (Bowman *et al*., 2003) and have been using similar formats in the assessment of food perception and preference for over a decade (e.g., Booth *et al.*, 2003).

The most effective way to elicit quantitative judgments of appetite is to arrange six to ten unlabelled points between two anchor phrases, chosen to encompass the contexts planned for those assessments of appetite. Additional unlabelled points should be provided beyond either or both anchors if the anchor is not the logical extreme (either 'not at all' or 'as much as I can imagine'). The points can be boxes, dashes, colons, dashes between colons, hatches on a line – all this is trivial. Indeed, an unmarked continuous line can be used if lack of structure within which to respond does not worry the assessors.

Instead of unnumbered lines or boxes, the points on the layout can be bare integers (including zero), perhaps fitted to a well known scoring system, such as marks out of five, seven or ten. If the assessors are familiar with percentages, they can be used, although the tens (or at most tens and fives) should be indicated as sufficient. Another advantage of this layout is that the rating is the score. Unnumbered boxes, dashes and fixed marks along a line are easily counted when scoring, whereas an unbroken line involves fuss with a ruler.

It is important that one of the two anchors is perceived and scored as zero, because the anchoring the numeral '1' imposes a ratio between the two anchors. If it is logically possible for the rating to go below zero (e.g., more satiated than sufficient to make me end the comparison meal), then negative integers can be used as well as the positive ones on the other side of that anchor. A score out of ten can be written or keyed, instead of being one of the options from an explicit number series. A percentage score should be inserted into a blank by the assessor.

Both zero and the top score, or a medium score, must of course be anchored by phrases appropriate to the situations tested in the study. The pair of anchoring words or phrases needs to be chosen at the same time as their positions on the layout for responses. An upper anchor of "as much as I can imagine", at one end of the response array, will work after a fashion in assessing ordinary influences on appetite. However, the quantitative judgments will be more precise if the upper anchor refers specifically to what is for the assessors a highly familiar and readily reconstructed point on the psychological scale being measured, e.g., ready to have lunch, eat a cookie or go for a drink or, more generally, ready to eat. Such ratings will not only be more realistic and accurate than line ratings from 'not at all' to 'extremely'. In addition, more of the response format will be used. Most importantly of all, there will be a familiar real situation for the test situation to be compared with.

Clearly then, there is no standard layout or wording for use to rate appetite in order to measure influences on it (or for rating on any other psychological scale). Hence the most important thing that an investigator should do, even if the layout was unfortunate in some way, is to report the exact wordings of the anchors and of the question answered by their use, and the score given to each anchor.

Basically, the format is irrelevant so long as only two anchors are used and end-artefacts are avoided, either by selecting samples/situations to keep responses well away from the end or by providing space for responses outside any anchor that is not at a logical extreme. For that reason, examples of layouts are not given in this paper. A much more important reason, however, is that the format of the ratings is irrelevant unless the study has been designed to produce explicable variations in appetite.

Multiple anchors. In logical principle, no more (and no fewer) than two anchor quantities should be used. Also they must be two levels of purely one concept – that is, rating must be unipolar, because 'opposites' are typically not the same concept. Just one point does not specify a line and so leaves the rater to choose the second anchor or, equivalently, the slope of the line (as 'magnitude estimation' attempts to do by asking for a second rating in ratio to the first).

Numerical rating has repeatedly demonstrated unequal spacing of the quantitative categories (e.g. "extremely", "slightly", "neither/nor") in one of the most widely used multiple-category formats, with nine anchors from "Like extremely" to

"Dislike extremely", misleadingly called the hedonic scale (there being no evidence that it measures pleasure as distinct from likelihood of choice).

Worse, these anchors use two distinct concepts: dislike is not the opposite of liking because the causes of the two reactions are usually completely different, e.g. a strange name or flavour or a taint, versus degrees of familiarity or slightly stronger or weaker than than the preferred level of flavouring. After 50 years, it is about time that users of that scale found out what is really happening, simply by using the categories of liking and dislike separately to obtain quantitative judgments and looking at the uncorrelated variance. One rating should be from "like extremely" to "neither like nor dislike" (with no intermediate anchors) and the other from dislike extremely to the zero dislike anchor.

The widely used 5 or 7 categories from strongly agree to strongly disagree have the same problem of unequal spacing. Therefore all the anchor phrases should be specified, along with their scoring, and not merely referred to as "Likert", "hedonic" or some other jargon.

Envoi

A rating needs two positions calibrated on observable quantities, such as zero and a familiar standard. With three or more anchor points, the slope of the psychophysical function may change at an intermediate point.

The score in the units of those two anchors does not imply any psychological scale, magnitude, process or state or what influence(s) operated on that rating. In order to measure motivation to eat or drink, or related sensations, emotions and perceptions, there has to be consistency among ratings on each concept or, far more informative, sensitivity of a rating to a specific influence on appetite.

References

Booth, D.A. (1990). How not to think about immediate dietary and postingestional influences on appetites and satieties. *Appetite* 14, 171-179.

Booth, D.A. (1991). Learned ingestive motivation and the pleasures of the palate. In R.C. Bolles (Ed.), *The hedonics of taste*, pp. 29-58. Hillsdale NJ: Erlbaum.

Booth, D.A. (1995). Cognitive processes in odorant mixture assessment. *Chemical Senses* 20, 639-643.

Booth, D.A. (2008). Physiological regulation through learnt control of appetites by contingencies among signals from external and internal environments. *Appetite* 51, 433-441.

Booth, D.A. (2009). Lines, dashed lines and "scale" ex-tricks. Objective measurements of appetite *versus* subjective tests of intake. *Appetite* 53, xxx-xxx [accompanying Short Communication].

Booth, D.A., & Freeman, R.P.J. (1993). Discriminative measurement of feature integration in object recognition. *Acta Psychologica* 84, 1-16.

Booth, D.A., Mather, P., & Fuller, J. (1982). Starch content of ordinary foods associatively conditions human appetite and satiation, indexed by intake and eating pleasantness of starch-paired flavours. *Appetite* 3, 163-184.

Booth, D.A., Mobini, S., Earl, T., & Wainwright, C.J. (2003). Market-optimum instrumental values from individual consumers' discriminations of standard sensory quality of the texture of short-dough biscuits. *Journal of Food Quality* 26(5), 425-439.

Booth, D.A., Thompson, A.L. & 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.G., & the UK Sjögren's Interest Group (2004). Measurement of fatigue and discomfort in primary Sjogren's syndrome using a new questionnaire tool. *Rheumatology* 43, 758-764.

Carroll, J.D., & Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling. *Psychometrika* 35, 283-xxx.

Conner, M.T., & Booth, D.A. (1988). Preferred sweetness of a lime drink and preference for sweet over non-sweet foods, related to sex and reported age and body weight. *Appetite* 10, 25-35.

Coombs, C. (1964). *A theory of data*. New York: John Wiley

Dibsdall, L.A., Wainwright, C.J., Read, N.W., & Booth, D.A. (1996). How fats and carbohydrates in familiar foods contribute to everyday satiety by their sensory and physiological actions. *British Food Journal* 99, 142-147.

Freeman, R.P.J., Richardson, N.J., Kendal-Reed, M.S., & Booth, D.A. (1993). Bases of a cognitive technology for food quality. *British Food Journal* 95 (9), 37-44.

Kissileff, H.R., Booth, D.A., Thornton, J.C., Pi-Sunyer, F.X., Pierson, R.N., & Lee, J. (2008). Human food intake is discriminatively sensitive to gastric signaling. *Appetite* 51, 759.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory*. 3rd Edition. New York: McGraw-Hill.

Rolls, B.J., Rolls, E.T., Rowe, E.A., & Sweeney, K. (1991). Sensory-specific satiety in man. *Physiology and Behavior* 27, 137-141.

Wyell, C.L., & Berridge, K.C. (2000). Intra-acccumbens amphetamine increases the conditioned incentive salience of sucrose reward. Enhancement of reward "wanting" without enhanced "liking" or response reinforcement. *Journal of Neuroscience* 20, 8122-8130.

Yeomans, M.R. (2000). Rating changes over the course of meals: what do they tell us about motivation to eat? *Neuroscience and Biobehavioral Reviews* 24, 249-259.