Chapter 6

Measurement of Sensory Thresholds

Abstract This chapter discusses the concept of threshold and contrasts the conceptual notion with the idea of threshold as a statistically derived quantity. A simple method for determining detection thresholds based on ASTM method E-679 is illustrated with a worked example. Other methods for determining thresholds are discussed as well as alternative analyses.

A light may be so weak as not sensibly to dispel the darkness, a sound so low as not to be heard, a contact so faint that we fail to notice it. In other words, a finite amount of the outward stimulus is required to produce any sensation of its presence at all. This is called by Fechner the law of the threshold—something must be stepped over before the object can gain entrance to the mind.

—(William James, 1913, p. 16)

Contents

6.1 Introduction: The Threshold Concept 125
6.2 Types of Thresholds: Definitions 127
6.3 Practical Methods: Ascending Forced Choice 128
6.4 Suggested Method for Taste/Odor/Flavor Detection Thresholds 129
6.4.1 Ascending Forced-Choice Method of Limits 129
6.4.2 Purpose of the Test 129
6.4.3 Preliminary Steps 130
6.4.4 Procedure 131
6.4.5 Data Analysis 131
6.4.6 Alternative Graphical Solution 131
6.4.7 Procedural Choices 133
6.5 Case Study/Worked Example 133
6.6 Other Forced Choice Methods 134
6.7 Probit Analysis 136
6.8 Sensory Adaptation, Sequential Effects, and Variability 136
6.9 Alternative Methods: Rated Difference, Adaptive Procedures, Scaling 137
6.9.1 Rated Difference from Control 137
6.9.2 Adaptive Procedures 138
6.9.3 Scaling as an Alternative Measure of Sensitivity 140
6.10 Dilution to Threshold Measures 140
6.10.1 Odor Units and Gas-Chromatography Olfactometry (GCO) 140
6.10.2 Scoville Units 142
6.11 Conclusions 142
Appendix: MTBE Threshold Data for Worked Example 143
References 145

6.1 Introduction: The Threshold Concept

One of the earliest characteristics of human sensory function to be measured was the absolute threshold. The absolute or detection threshold was seen as an energy level below which no sensation would be produced by a stimulus and above which a sensation would reach consciousness. The concept of threshold was central to Fechner's psychophysics. His integration of Weber's law produced the first psychophysical relationship. It depended upon the physical intensity being measured with the threshold for sensing changes as the unit (Boring, 1942). Early physiologists like Weber and Fechner would use the classical method of limits to measure this point of discontinuity, the beginning of the psychophysical function. In the method of limits, the energy level would be raised and lowered and the average point at which the observer changed response from "no sensation"

H.T. Lawless, H. Heymann, *Sensory Evaluation of Food*, Food Science Text Series,
DOI 10.1007/978-1-4419-6488-5_6, © Springer Science+Business Media, LLC 2010

to "yes, I perceive something" would be taken as the threshold. This specification of the minimum energy level required for perception was one of the first operating characteristics of sensory function to be quantified. Historically, the other common property to be measured was the difference threshold or minimal increase in energy needed to produce a noticeable increase in sensation. Together, these two measures were used to specify the psychophysical function, which to Fechner was a process of adding up difference thresholds once the absolute (minimum) threshold had been surpassed.

In practice, some complications arise in trying to apply the threshold idea. First, anyone who attempts to measure a threshold finds that there is variability in the point at which observers change their response. Over multiple measurements there is variability even within a single individual. In a sequence of trials, even within the same experimental session, the point at which a person changes his or her responses will differ. An old story has it that S.S. Stevens, one of the pioneers of twentieth century psychophysics, used the following classroom demonstration at Harvard: Students were asked to take off their wristwatches and hold them at about arm's length, then count the number of ticks they heard in 30 s (back in the day when spring-wound watches still made ticking sounds). Assuming the watch of one of these Harvard gentlemen made uniform ticking sounds, the common result that one would hear some but not all of the ticks illustrated the moment-to-moment variation in auditory sensitivity. Of course, there are also differences among individuals, especially in taste and smell sensitivity. This led to the establishment of common rules of thumb for defining a threshold, such as the level at which detection occurs 50% of the time.

The empirical threshold (i.e., what is actually measured) remains an appealing and useful concept to many workers involved in sensory assessments. One example is in the determination of flavor chemicals that may contribute to the aromatic properties of a natural product. Given a product like apple juice, many hundreds of chemical compounds can be measured through chemical analysis. Which ones are likely to contribute to the perceived aroma? A popular approach in flavor analysis assumes that only those compounds that are present in concentrations above their thresholds will contribute. A second example of the usefulness of a threshold is in defining a threshold for taints or off-flavors in a product. Such a value has immediate practical implications for what may be acceptable versus unacceptable levels of undesired flavor components. Turning from the product to the sensory panelists, a third application of thresholds is as one means of screening individuals for their sensitivity to key flavor components. The measurement of a person's sensitivity has a long history in clinical testing. Common vision and hearing examinations include some measurements of thresholds. In the chemical senses, threshold measurements can be especially useful, due to individual differences in taste and smell acuity. Conditions such as specific anosmia, a selective olfactory deficit, can be important in determining who is qualified for sensory test panel participation (Amoore, 1971).

Another appealing aspect of the threshold concept is that the values for threshold are specified in physical intensity units, e.g., moles per liter of a given compound in a product. Thus many researchers feel comfortable with threshold specification since it appears to be free from the subjective units of rating scales or sensory scores. However, threshold measurements are no more reliable or accurate than other sensory techniques and are usually very labor intensive to measure. Perhaps most importantly, thresholds represent only one point on a dose–response curve or psychophysical function, so they tell us little about the dynamic characteristics of sensory response as a function of changes in physical concentration. How the sensory system behaves above threshold requires other kinds of measurements.

In this chapter, we will look at some threshold definitions and approaches and their associated problems. Next, we will examine some practical techniques for threshold measurement and discuss a few applications. Throughout, we will pay special attention to the problems of variability in measurement and the challenges that this poses for researchers who would use thresholds as practical measures of peoples' sensitivities to a given stimulus, or conversely, of the potency or biological activity of that stimulus in activating sensory perceptions. Most of the examples chosen come from olfaction and taste, as the chemical senses are especially variable and are prone to difficulties due to factors such as sensory adaptation.

6.2 Types of Thresholds: Definitions

What is a threshold? The American Society for Testing and Materials (ASTM) provides the following definition that captures the essence of the threshold concept for the chemical senses: "A concentration range exists below which the odor or taste of a substance will not be detectable under any practical circumstances, and above which individuals with a normal sense of smell or taste would readily detect the presence of the substance."—ASTM method E-679-79 (2008a, p. 36).

Conceptually, the absolute or detection threshold is the lowest physical energy level of a stimulus or lowest concentration in the case of a chemical stimulus that is perceivable. This contrasts with empirical definitions of threshold. When we try to measure this quantity, we end up establishing some practical rule to find an arbitrary value on a range of physical intensity levels that describes a probability function for detection. In 1908, the psychologist Urban recognized the probabilistic nature of detection and called such a function a psychometric function, as shown in Fig. 6.1 (Boring, 1942). We portray this function as a smooth curve in order to show how the original concept of a fixed threshold boundary was impossible to measure in practice. That is, there is no one energy level below which detection never occurs and above which detection always occurs.

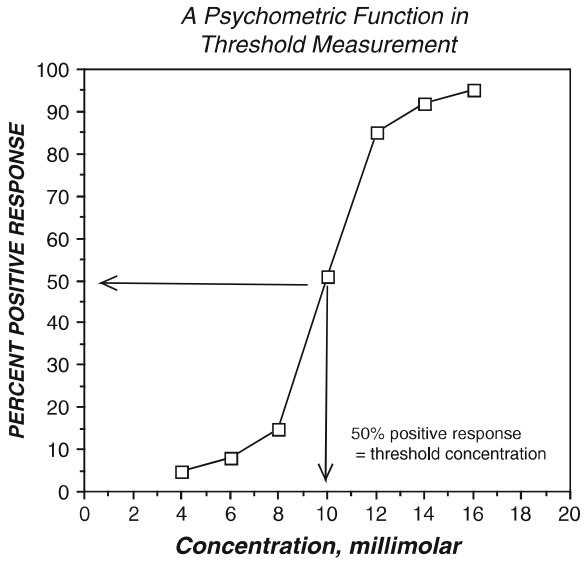

Fig. 6.1 A psychometric function.

It is not a sudden step function. There is a probability function determined by an empirical method of measurement on which we define some arbitrary point as the threshold.

Recognition thresholds are also sometimes measured. These are the minimum levels that take on the characteristic taste or smell of the stimulus and are often a bit higher than detection thresholds. For example, dilute NaCl is not always salty, but at low concentrations just above the detection threshold is perceived as sweet (Bartoshuk et al., 1978). The concentration at which a salty taste is apparent from NaCl is somewhat higher. In food research, it is obvious that the recognition threshold for a given flavor in a food would be a useful thing to know, and perhaps more useful than detection thresholds, since both the percept and the appropriate label have been made consciously available and actionable to the taster. In the case of off-flavors or taints, recognition may have strong hedonic correlates in predicting consumer rejection.

To be recognized and identified, discrimination from the diluent is only one requirement. In addition, the observer must assign the appropriate descriptor word to the stimulus. However, it is difficult to set up a forced-choice experiment for identification in some modalities. In taste, for example, you can have the observer pick from the four (or five) taste qualities, but there is no assurance that these labels are sufficient to describe all sapid substances (O'Mahony and Ishii, 1986). Furthermore, one does not know if there is an equal response bias across all four alternatives. Thus the expected frequencies or null hypothesis for statistical testing or difference from chance responding is unclear. In an experiment on bitter tastes, Lawless (1980) attempted to control for this bias by embedding the to-be-recognized bitter substances in a series that also included salt, acid, and sugar. However, the success of such a procedure in controlling response biases is unclear and at this time there are no established methods for recognition thresholds that have adequately addressed this problem.

The difference threshold has long been part of classical psychophysics (see Chapter 2). It represents the minimum physical change necessary in order for a person to sense the change 50% of the time. Traditionally, it was measured by the method of constant stimuli (a method of comparison to a constant reference) in

which a series of products were raised and lowered around the level of the reference. The subject would be asked to say which of member of the pair was stronger and the point at which the "stronger" judgment occurred 75% (or 25%) of the time was taken as the difference threshold or "just-noticeable-difference" (JND).

One can think of sensory discrimination tests (triangles and such) as a kind of difference threshold measurement. The main difference between a psychophysical threshold test and a sensory discrimination test is that the psychophysical procedure uses a series of carefully controlled and usually simple stimuli of known composition. The sensory product test is more likely to have only two products, and the pair is either deemed different or not, based on a criterion of statistical significance. But clearly the two kinds of tests are related. Along these lines, one can think of the absolute threshold as a special case of a difference threshold, when the standard happens to be some blank or baseline stimulus (such as pure air or pure water).

In addition to detection, recognition, and difference thresholds, a fourth category is the terminal threshold or region in which no further increase in response is noted from increasing physical stimulus intensity (Brown et al., 1978). In other words, the sensory response has reached some saturation level, beyond which no further stimulation is possible due to maximal responding of receptors or nerves or some physical process limiting access of the stimulus to receptors. This makes sense in terms of neurophysiology as well. There are only a limited number of receptors and nerves and these have a maximal response rate. This idea fits well with the notion of a threshold as a discontinuity or inflection point in the psychophysical function (Marin et al., 1991).

However, in practice, this level is rarely approached. There are few foods or products in which the saturation level is a common level of sensation, although some very sweet confections and some very hot pepper sauces may be exceptions. For many continua, the saturation level is obscured by the addition of new sensations such as pain or irritation (James, 1913). For example, some odors have a down-turn in the psychophysical function at high levels, as trigeminal irritation begins to take place, that may in turn have an inhibiting effect on odor intensity (Cain, 1976; Cain and Murphy, 1980). Another example is in the bitter side taste of saccharin. At high levels, the bitterness will overtake the sweet sensation for some individuals. This makes it difficult to find a sweetness match of saccharin to other sweeteners at high levels (Ayya and Lawless, 1992). Further increases in concentration only increase bitterness and this additional sensation has an inhibiting effect on sweet perception. So although saturation of response seems physiologically reasonable, the complex sensations evoked by very strong stimuli mediate against any measurement of this effect in isolation.

Recently, a new type of threshold has been proposed for consumer rejection of a taint or off-flavor. Prescott et al. (2005) examined the levels at which consumers would show an aversion to cork taint from trichloroanisole in wines. Using a paired preference tests with increasing levels of trichloroanisole, they defined the rejection threshold as the concentration at which there was a statistically significant preference for an untainted sample. This novel idea may find wide application in flavor science and in the study of specific commodities (like water) in which the chemistry and origins of taints are fairly well understood (for another example, see Saliba et al., 2009). The method in its original form requires some refinement as to the criterion for threshold because statistical significance is a poor choice. As they noted in their paper, the level of statistical significance depends upon the number of judges, a more or less arbitrary choice of the experimenter (not a function of the sensory response of the participants). A better choice would be something akin to the difference threshold, i.e., the concentration at which 75% preference was reached. Of course, confidence intervals can be drawn around any such level for those that need statistical assurance.

6.3 Practical Methods: Ascending Forced Choice

In the early days of psychophysics, the method of limits was the most common approach to measuring thresholds. In this procedure, stimulus intensity would be raised in an ascending series and then lowered in a descending series to find points at which the observer's response changed from a negative to a positive response or from positive to negative. Over several

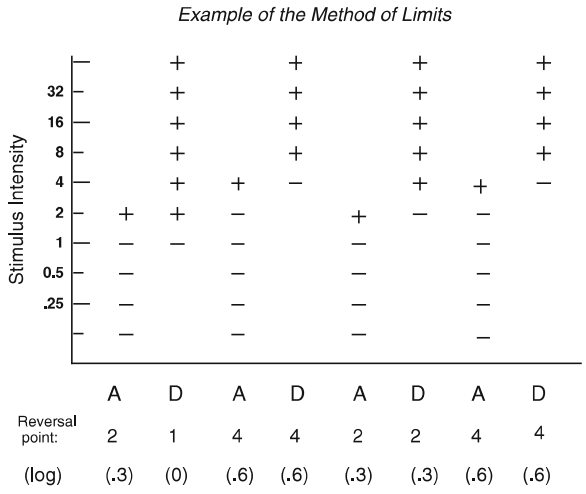

Fig. 6.2 Method of limits example.

ascending and descending runs, an average changing point could be taken as the best estimate of threshold (McBurney and Collings, 1977). This method is illustrated in Fig. 6.2.

Although this procedure seems straightforward, it has several problems. First, the descending series may cause such fatigue or sensory adaptation that the observer fails to detect stimulus presentations that would be clearly perceived if they were presented in isolation. To avoid the adaptation or fatigue problem that is common in the taste and smell senses, the method is usually performed only in an ascending series. A second difficulty is that different persons may set different criteria for how much of a sensation they require before changing their response. Some people might be very conservative and have to be positively sure before they respond, while others might take any inkling at all as a reason to report a sensation. Thus the classical method of limits is contaminated by the panelist's individual bias or criterion, which is not a function of their sensitivity, i.e., what the test is actually trying to measure. This is a central issue in the theory of signal detection (see Chapter 5). To address the problem of uncontrolled individual criterion, later workers introduced a forced choice element to the trials at each intensity level or concentration step (e.g., Dravnieks and Prokop, 1975). This combines the method of limits with a discrimination test. The task requires that the observer gives objective proof of detection by discriminating the target stimulus from the background level. A forced choice technique is compatible with signal detection principles and is bias free, since the observer does not choose whether or not to respond—response is required on each trial.

6.4 Suggested Method for Taste/Odor/Flavor Detection Thresholds

6.4.1 *Ascending Forced-Choice Method of Limits*

This procedure is based on a standard method designated ASTM E-679 (ASTM, 2008a). It follows the classical method of limits, in which the stimulus intensity, in this case concentration of a taste or odor chemical, is raised in specified steps until the substance is detected. The procedure adds a forced choice task in which the substance to be detected is embedded in a set of stimuli or products that includes other samples that do not contain any of the added substance. The stimulus or product with the taste or odor chemical is called a "target" and the other items with no added chemical are often referred to as "blanks." One can use various combinations of targets and blanks, but it is common to have one target and in the case of E-679, two additional blanks. So the task is a three-alternative forced choice task (3-AFC), because the person being tested is forced to choose the one different sample in the set of three. That is if they are uncertain, they are told to guess.

6.4.2 *Purpose of the Test*

This method is designed to find the minimum level (minimum concentration) of a substance that is detected by 50% of the sample group. In practice, this is calculated as the geometric mean of the individual threshold estimates. The geometric mean is a reasonable choice because it is often very close to the median (50th percentile) of a positively skewed distribution. Threshold data tend to show high outliers, i.e., some insensitive individuals cause positive skew.

6.4.3 Preliminary Steps

Before the test is conducted, there are several tasks and some choices that must be made, as shown in Table 6.2. First, a sample of the substance of known purity must be obtained. Second, the diluent (solvent, base) or carrier must be chosen. For the detection threshold for flavors, for example, it is common to use some kind of pure water such as deionized or distilled. Third, the size of the concentration steps must be chosen. It is common to use factors of two or three. In other words, the concentrations will be made up in a geometric progression, which are equal steps on a log scale. Fourth, some sample concentrations should be set up for preliminary or "benchtop screening" to estimate the range in which the threshold is likely to occur. This can be done by successive dilutions using factors of five or ten, but beware the effects of adaptation on reducing one's sensitivity to subsequent test items. Exposure to a strong sample early in this series may cause subsequent samples to seem odorless or tasteless, when they might in fact be perceived when tasted alone. The outcome of the preliminary test should bracket the likely concentration range, so that most, if not all, of the people who participate in the formal test will find an individual threshold estimate somewhere within the test series. It is common to use about eight to ten steps in this procedure.

Next, the panel should be recruited or selected. A sample group should have at least 25 participants. If the goal is to generalize the result to some larger population, then the panel should be representative of that population with respect to age, gender, and so on and a larger panel of 100 or more is recommended. It is common practice to exclude people with known health problems that could affect their sense of taste or smell and individuals with obvious sensory deficits in the modality being tested. Of course, all the appropriate setup work must be done that is associated with conducting any sensory test, such as securing a test room that is free from odors and distractions, scheduling the panelists, setting up the questionnaire or answer sheet, writing instructions for the participants. See Chapter 3 for further details on good practices in sensory testing. For threshold work it is especially important to have clean odor-free glassware or plastic cups that are absolutely free of any odor that would contaminate the test samples. In odor testing the sample vessels are usually covered to preserve the equilibrium in the headspace above the liquid. The covers are removed by each panelist at the moment of sniffing and then replaced. Finally, external sources of odor must be minimized or eliminated, such as use of perfumes or fragrances by participants, hand lotions, or other fragranced cosmetics that could contaminate the sample vessels or the general area. Avoid using any markers or writing instruments that might have an odor when marking the samples. As always, sample cups or vessels should be marked with blind codes, such as randomly chosen three digit numbers. The experimenter must set up random orders for the three items at each step and use a different randomization for each test subject. This

Table 6.1 Types of thresholds

Detection (absolute) threshold:	Point at which the substance is differentiated from the background
Recognition threshold:	Point at which the substance is correctly named
Difference threshold:	(just-noticeable-difference, JND) Point at which the change in concentration is noted
Terminal threshold:	Point at which no further intensity increase is found with increasing concentration
Consumer rejection threshold:	Point at which a consumer preference occurs for a sample *not* containing the substance

Table 6.2 Preliminary tasks before threshold testing

1. Obtain test compound of known purity (note source and lot number)
2. Choose and obtain the solvent, carrier, or food/beverage system
3. Set concentration/dilution steps, e.g., 1/3, 1/9, 1/27
4. Begin benchtop screening to bracket/approximate threshold range
5. Choose number of dilution steps
6. Recruit/screen panelists. $N \geq 25$ is desirable
7. Establish procedure and pilot test if possible
8. Write verbatim instructions for panelists

should be recorded on a master coding sheet showing the randomized three-digit codes and which sample is the correct choice or target item.

6.4.4 Procedure

The steps in the test are shown in Table 6.3. The participant or test subject is typically seated before a sample tray containing the eight or so rows of three samples. Each row contains one target sample and two blank samples, randomized. The instructions, according to E-679-04 (ASTM, 2008a) are the same as in the triangle test, that is to pick out the sample which is different from the other two. The subject is told to evaluate the three samples in each row once, working from left to right. The test proceeds through all the steps of the concentration series and the answers from the subject are recorded, with a forced guess if the person is uncertain. According to E-679, if a person misses at the highest level, that level will be repeated. If a person answers correctly through the entire series, the lowest level will also be repeated for confirmation. If the response changes in either case, it is the repeated trial that is counted.

6.4.5 Data Analysis

Figure 6.3 shows an example of how the data are analyzed and the threshold value is determined. First, an individual estimated threshold is determined for each person. This is defined as the concentration that is the geometric mean of two values (the square root of the product of the two values). One value is the concentration at which they first answered correctly and all higher concentrations were also correct. The other value is the concentration just below that, i.e., the last incorrect judgment. This interpolation provides some protection against the fact that the forced-choice procedure will tend to slightly overestimate the individual's threshold (i.e., the concentration at which they have a 0.5 probability of sensing that something is different from the blanks). If the subject gets to the top of the series with an incorrect judgment, or starts at the bottom with all judgments correct, then a value is extrapolated beyond the test series. At the top, it is the geometric mean of the highest concentration tested and the next concentration that would have been used in the series if the series had been continued. At the bottom, it is the geometric mean of the lowest concentration tested and the next lower concentration that would have been used had the series been continued lower. This is an arbitrary rule, but it is not unreasonable. Once these individual best estimates are tabulated, the group threshold is the geometric mean of the individual values. The geometric mean is easily calculated by taking the log of each of the individual concentration values, finding the average of the logs, and then taking the antilog of this value (equivalent to taking the Nth root of the product of N observations).

6.4.6 Alternative Graphical Solution

An alternative analysis is also appropriate for this kind of data set. Suppose that 3-AFC tests had been conducted and the group percent correct calculated at each step. For examples, see Antinone et al. (1994)

Table 6.3 Ascending forced-choice testing steps

1.	Obtain randomized or counterbalanced orders via software program or random number generator.
2.	Setup trays or other staging arrangements for each participant, based on random orders.
3.	Instruct participants in procedure per verbatim script developed earlier.
4.	Show suprathreshold example (optional).
5.	Present samples and record results. Force a choice if participant is unsure.
6.	Tally results for panel as series of correct/incorrect answers.
7.	Calculate estimated individual thresholds: Geometric mean of first correct answer with all higher concentrations correct and last incorrect step.
8.	Take geometric mean of all individual threshold estimates to get group threshold value.
9.	Plot graphic results of proportion correct against log concentration. Interpolate 66.6% correct point and drop line to concentration axis to get another estimate of threshold (optional).
10.	Plot upper and lower confidence interval envelopes based on $\pm 1.96(p\ (1-p)/N)$. Drop lines from the upper and lower envelopes at 66.6% to concentration axis to convert envelope to concentration interval.

	Concentration ug/L									
Panelist	2	3.5	6	10	18	30	60	100	BET	log(BET)
1	+	o	+	+	+	+	+	+	2.6	0.415
2	o	o	o	+	+	+	+	+	7.7	0.886
3	o	+	o	o	o	o	+	+	42	1.623
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
N	o	o	+	+	o	+	o	+	77	1.886
Prop. Corr.	0.44	0.49	0.61	0.58	0.65	0.77	0.89	0.86	Mean (log(BET))	1.149
									10^1.149 =	14.093

Fig. 6.3 Sample data analysis from ascending 3-AFC method Notes: Correct choices indicated by + and incorrect by o. BET, Best estimate of individual threshold, defined as the geometric mean of the first correct trial with all subsequent trials correct and the previous (incorrect) trial. The group threshold is calculated from the geometric mean of the BET values. In practice, this is done by taking the logs of the BET values, finding the mean of the logs (x), then taking the antilog of that value (or 10^x).

and Tuorila et al. (1981). We can take the marginal count of the number of correct choices from the bottom row in Fig. 6.3 and expressing it as the proportion correct. As the concentration increases, this proportion should go from near the chance level (1/3) to nearly 100% correct. Often this curve will form an S-curve similar to the cumulative normal distribution. The threshold can then be defined as the level at which performance is 50% correct, once we have adjusted the data for chance, i.e., the probability that a person could guess correctly (Morrison, 1978; Tuorila et al., 1981; Viswanathan et al., 1983). This is done by Abbott's formula, a well-known correction for guessing as shown in Eqs. (6.1) and (6.2):

$$P_{\text{corr}} = (P_{\text{obs}} - P_{\text{chance}})/(1 - P_{\text{chance}}) \qquad (6.1)$$

where P_{corr} is the chance-corrected proportion, P_{obs} is the observed proportion correct in the data, and P_{chance} is the chance probability, e.g., 1/3 for the 3-AFC. Another form is

$$P_{\text{req}} = (P_{\text{chance}} - P_{\text{corr}})/(1 - P_{\text{chance}}) \qquad (6.2)$$

where P_{req} is the observed proportion that is required in order to achieve a certain chance corrected level of performance. So if one needed to get a chance corrected proportion of 0.5 (i.e., a threshold, 50% detection) in a 3-AFC test, you would need to see 1/3 + 0.5 (1–1/3) or 2/3 (= 66.7%) correct.

Once a line or curve is fitted to the data, the concentration at which the group would achieve 66.6% correct can be solved (or simply interpolated by eye if the data are fairly linear and a curve is fit by eye). A useful equation that can be fit to many data sets is based on logistic regression, shown in Eq. (6.3) (e.g., Walker et al., 2003).

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 \log C \qquad (6.3)$$

where p is the proportion correct at concentration C and b_0 and b_1 are the intercept and slope. The quantity, $p/1–p$, is sometimes referred to as the odds ratio. The interpolation is shown in Fig. 6.4. Note that this also

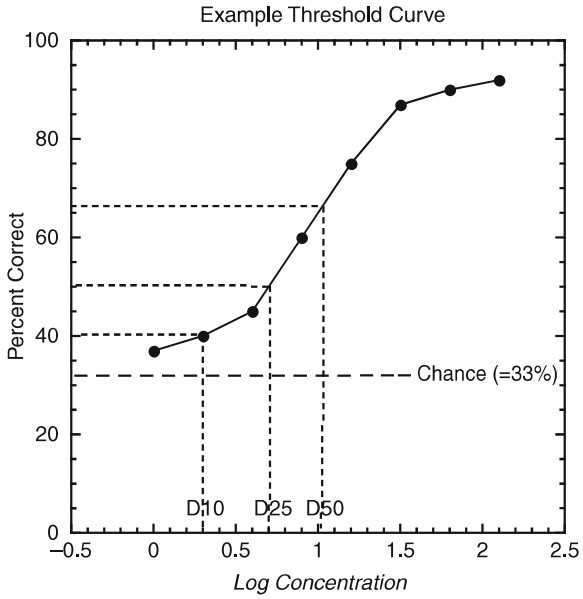

Fig. 6.4 Sample threshold curve and interpolation. D10, D25 and D50 show interpolated detection levels for 10%, 25% and 50% of persons, respectively.

allows one to estimate percentages of the population that would detect with other probabilities and not just the arbitrary 50% detection that we use as the threshold value. That is, one could interpolate at 10 or 90% detection if that was of interest. A lower percentage of detection might be of interest, for example, in setting level to protect consumers from an off-flavor or taint.

This graphical method has certain assumptions and limitations that the user should be aware of. First, it assumes that persons are either detecting or guessing (Morrison, 1978). In reality, every person has an individual threshold gradient or gradually increasing probability of detection around their own threshold. Second, the model does not specify what *percent of the time* that a given percentage of the group will detect. In the data set examined below, the ASTM method and the graphical solution provide a good estimate of when 50% of the group will detect 50% of the time. More extensive statistical models have been developed for this kind of data and an extensive paper on alternative statistical analyses is given in USEPA (2001), again using the data set we have chosen as an example below.

6.4.7 Procedural Choices

Note that although the instructions are the same as in the triangle test, all the possible combinations of the three samples are not used, i.e., the test is not a fully counterbalanced triangle. Only the three possible orders that are given by combinations of two blanks and one target are used. In a fully counterbalanced triangle, the additional three combinations of two targets and one blank would have been used (thus a total of six possible), but this is not done according to E-679. For taste or flavor, there is generally no rinsing between samples, although testers may be instructed to rinse between rows (triads). If possible, it is wise to give the subject a preliminary sample at a detectable level, in order to show them the target item that they will be trying to sense in the test. Of course, one must be careful when using such an above-threshold sample so that it does not adapt or fatigue the senses. An appropriate waiting time and/or rinsing requirement should be used to prevent any effect on the subsequent test samples in the formal test. The experimenters should also decide whether they will allow re-tasting or not. Re-tasting could either confuse the subjects or it might help them get a better idea of which item is the target. We would generally argue against re-tasting, because that will introduce a variable that is left up to the individual subject and will thus differ among people. Some will choose to re-taste and others will not. So, on the basis of maintaining a consistent test procedure across all participants, re-tasting is generally not recommended.

Another important choice is that of a "stopping rule." In the published version of E-679, every subject must continue to the top of the series. There are some pitfalls in this, because of the possibility that the senses will become fatigued or adapted by the high levels at the top of the series, especially for an individual with a low personal threshold. For this reason, some threshold procedures introduce a "stopping rule." For example, the panelist may be allowed to stop tasting after giving three correct answers at adjacent levels (Dravnieks and Prokop, 1975). This prevents the problem of exposing a sensitive individual to an overwhelming stimulus at high levels. Such an experience, if unpleasant (such as a bitter taste), might even cause them to quit the test. On the downside, the introduction of a stopping rule can raise the false positive rate. We can think of a false positive as finding a threshold value for an individual that is due to guessing only. In the most extreme case, it would be a person who is completely insensitive (e.g., anosmic to that compound if it is an odor threshold) finding a threshold somewhere in the series. With an eight-step series, for the ASTM standard rule (everyone completes the series), the probability of finding a threshold somewhere in steps one through eight, for a completely anosmic person who is always guessing is 33.3%. For the three-in-a-row stopping rule, the chances of the anosmic person making three lucky guesses in a row somewhere rise above 50%. The sensory professional must weigh the possible negatives from exposing the participant to strong stimuli against the increased possibility of false positives creating a low-threshold estimate when using a stopping rule.

6.5 Case Study/Worked Example

For the ascending forced choice method of limits (ASTM E-679), we can use a published data set for odor thresholds. The actual data set is reproduced in

the Appendix at the end of this chapter. The data are from a study conducted to find the odor detection threshold for methyl tertiary butyl ether (MTBE), a gasoline additive that can contaminate ground water, rendering some well waters unpotable (Stocking et al., 2001; USEPA, 2001). The ASTM procedure was followed closely, including the triangle test instructions (choose the sample different from the other two), using the 3-AFC in eight concentration steps differing by a factor of about 1.8. Individual best estimates were taken as the geometric mean of the last step missed and the first step answered correctly, with all higher steps also correct. Individuals who got the first and all subsequent steps correct (there were 10/57 or 17.5% of the group) had their estimated threshold assigned as the geometric mean of the first concentration and the hypothetical concentration one step below that which would have been used had the series been extended down. A similar extrapolation/estimation was performed at the high end for persons that missed the target on the eighth (highest) level.

The geometric mean of the individual threshold estimates across a panel of 57 individuals, balanced for gender and representing a range of ages, was 14 μg/l (14 ppb). Figure 6.5 shows the graphical solution, which gives a threshold of about 14 ppb, in good agreement with the geometric mean calculation. This is the interpolated value for 66.7% correct, the chance-adjusted level for 50% probability of detection in the group. Confidence intervals (CI) for this level can be found by constructing upper and lower curves form an envelope of uncertainty around the fitted curve. The standard error is given by the square root of $(p(1–p)/N)$ or in this case 0.062 for $p = 1/3$ and $N = 57$.

The 95% CI is found by multiplying the z-score for 0.95 (= 1.96) times the standard error, in this case equal to ± 0.062(1.96) or ± 0.122. Constructing curves higher and lower than the observed proportions by this amount will then permit interpolation at the 66.7% level to find concentrations for the upper and lower CI bounds. This method is simple, but it provides conservative (wider) estimate of the confidence intervals that found with some other statistical methods such as bootstrap analysis (USEPA, 2001). Another method for error estimation based on the standard error of the regression line is given in Lawless (2010).

Note that by the graphical method, the interpolated value for 10% detection (= 40% correct by Abbott's formula) will be at about 1–2 ppb. Similarly the interpolated value for 25% detection (50% correct by Abbott's formula), will be between 3 and 4 ppb. These values are practically useful to a water company who wanted to set lower limits on the amount of MTBE that could be detected by proportions of the population below the arbitrary threshold value of 50% (Dale et al., 1997).

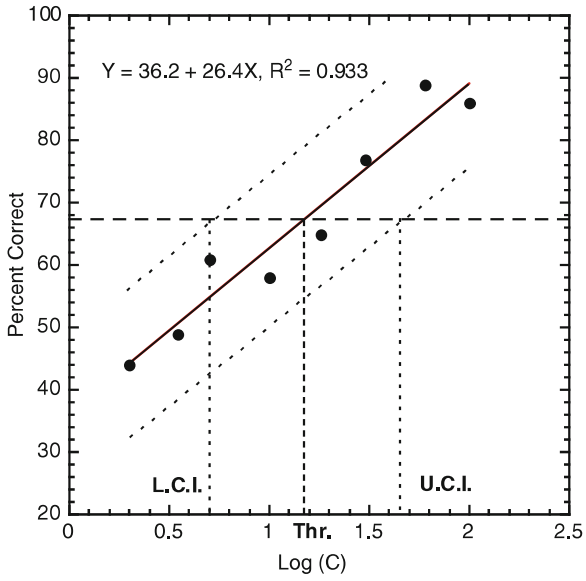

Fig. 6.5 Interpolation of threshold from the data of Stocking et al. (2001).

6.6 Other Forced Choice Methods

Ascending forced-choice procedures are widely used techniques for threshold measurement in the experimental literature on taste and smell. One early example of this approach is in the method for determining sensitivity to the bitter compound phenylthiourea, formerly called phenylthiocarbamide or PTC, and the related compounds 6-*n*-propylthiouracil or PROP. Approximately one-third of Caucasian peoples are insensitive to the bitterness of these compounds, as a function of several mutations in a bitter receptor that usually manifests as a simple homozygous recessive status for this trait (Blakeslee, 1932; Bufe et al., 2005). Early researchers felt the need to have a very stringent test of threshold, so they intermingled four blank samples (often tap water) with four target samples at each

concentration step (Harris and Kalmus, 1949). The chance probability of sorting correctly is only 0.014, so this is a fairly difficult test. In general, the formula for the chance probability of sorting any one level of X target samples among N total samples is given by Eq. (6.4):

$$p = X!/[N!/(N - X)!] \tag{6.4}$$

Obviously, the larger the number of target and blank samples, the more stringent the test and the higher the eventual threshold estimate. However, arbitrarily increasing X and N may make the task tedious and may lead to other problems such as fatigue and flagging motivation among the participants. The rigor of the test estimate must be weighed against undue complexity that could lead to failures to complete the series or poor quality data.

Another example of a threshold test for olfaction is Amoore's technique for assessing specific anosmia (Amoore, 1979; Amoore et al., 1968). Specific anosmia describes a deficit in the ability to smell a compound or closely related family of compounds among people with otherwise normal olfactory acuity. Being classified as anosmic was operationally defined by Amoore as having olfactory detection thresholds more than two standard deviations above the population mean (Amoore et al., 1968). The test is sometimes called a "two-out-of-five" test because at each concentration level there are two target stimuli containing the odorant to be tested and three diluent or blank control samples. The tester must sort the samples correctly in this two-out-of-five test, and the chance probability of obtaining correct sorting by merely guessing is one in ten. Performance is normally confirmed by testing the next highest concentration (an example of a "stopping rule"). The chance occurrence of sorting correctly on two adjacent levels is then 1 in 100. This makes the test somewhat difficult but provides a good deal of insurance against a correct answer by guessing.

Another way to reduce the chance performance on any one level is to require multiple correct answers at any given concentration. This is part of the rationale behind the Guadagni multiple pairs test (Brown et al., 1978) in which up to four pairs may be given for a two-alternative forced choice test in quadruplicate at any one concentration. Brown et al. commented upon the user-friendliness of this technique, i.e., how simple it was to understand and administer to participants. A variation was used by Stevens et al. (1988) in a landmark paper on the individual variability in olfactory thresholds. In this case, five correct pairs were required to score the concentration as correctly detected, and this performance was confirmed at the next highest concentration level. The most striking finding of this study was that among the three individuals tested 20 times, their individual thresholds for butanol, pyridine, and phenylethylmethylethyl carbinol (a rose odorant) varied over 2,000- to 10,000-fold in concentration. Variation within an individual was as wide as the variation typically seen across a population of test subjects. This surprising result suggests that *day-to-day variation in olfactory sensitivity is large and that thresholds for an individual are not very stable* (for an example, see Lawless et al., 1995). More recent work using extensive testing of individuals at each concentration step suggests that these estimates of variability may be high. Walker et al. (2003) used a simple yes/no procedure (like the A, not-A test, or signal detection test) with 15 trials of targets and 15 trials of blanks at each concentration level. Using a model for statistical significant differences between blank and target trials, they were able to get sharp gradients for the individual threshold estimates.

In summary, an ascending forced-choice method is a reasonably useful compromise between the need to precisely define a threshold level and the problems encountered in sensory adaptation and observer fatigue when extensive measurements are made. However, the user of an ascending forced-choice procedure should be aware of the procedural choices that can affect the obtained threshold value. The following choices will affect the measured value: the number of alternatives (both targets and blanks), the stopping rule, or the number of correct steps in a row required to establish a threshold, the number of replicated correct trials required at any one step, and the rule to determine at what level of concentration steps the threshold value is assigned. For example, the individual threshold might be assigned at the lowest level correct, the geometric mean between the lowest level correct and highest level incorrect. Other specific factors include the chosen step size of concentration units (factors of two or three are common in taste and smell), the method of averaging or combining replicated ascending runs on the same individual and finally the method of averaging or combining group data. Geometric means are commonly used for the last two purposes.

6.7 Probit Analysis

It is often useful to apply some kind of transformation or graphing method to the group data to linearize the curve used to find the 50% point in a group. Both the psychometric curve that represents the behavior of an individual in multiple trials of a threshold test and the cumulative distribution of a group will resemble an S-shaped function similar to the cumulative normal distribution. A number of methods for graphing such data are shown in the ASTM standard E-1432 (ASTM, 2008b). One simple way to graph the data is simply to plot the cumulative percentages on "probability paper." This pre-printed solution provides a graph in which equal standard deviations are marked off along the ordinate, effectively stretching the percentile intervals at the ends and compressing them in the midrange to conform to the density of the normal distribution. Another way to achieve the straightening of the S-shaped response curve is to transform the data by taking *z*-scores. Statistical packages for data analysis often provide options for transformation of the data.

A related method was once widely used in threshold measurement, called Probit analysis (ASTM, 2008b; Dravnieks and Prokop, 1975; Finney, 1971). In this approach, the individual points are transformed relative to the mean value, divided by the standard deviation and then a constant value of +5 is added to translate all the numbers to positive values for convenience. A linear fitted function can now be interpolated at the value of 5 to estimate the threshold as in Fig. 6.6. The conversion (to a *z*-score +5) tends to make an S-shaped curve more linear. An example of this can be found in the paper by Brown et al. (1978), using data from a multiple paired test. First the percent correct is adjusted for chance. Then the data are transformed from the percent correct (across the group) at each concentration level by conversion to *z*-scores and a constant of 5 is added. The mean value or Probit equal to 5 can be found by interpolation or curve fitting. An example of this technique for estimating threshold from a group of 20 panelists is shown in Meilgaard et al. (1991) and in ASTM (2008b). Probit plots can be used for any cumulative proportions, as well as ranked data and analysis of individuals who are more extensively tested than in the 3-AFC method example shown earlier.

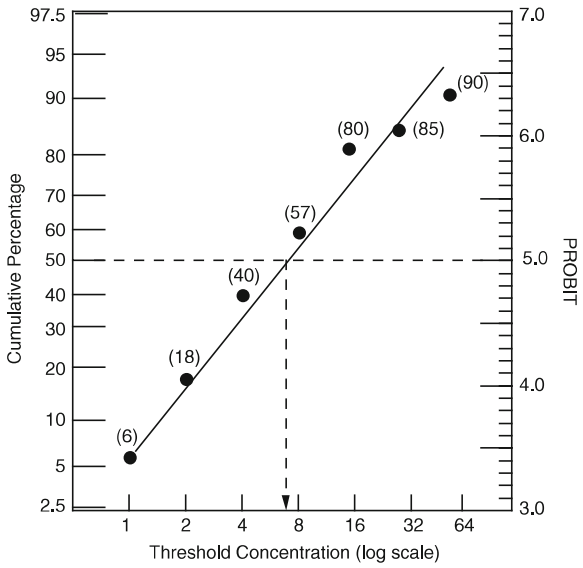

Fig. 6.6 An example of Probit analysis. Numbers in parentheses are the cumulative percentages of panelists reaching threshold at each concentration step. Note the uneven scale on the left axis. Probits mark off equal standard deviations and are based on the *z*-score for any proportion plus the constant, 5. Interpolation at the 50% or Probit 5.0 gives the threshold value.

6.8 Sensory Adaptation, Sequential Effects, and Variability

Individual variability, both among a group of people and within an individual over repeated measurements presents a challenge to the idea that the threshold is anything like a fixed value. For example, stable olfactory thresholds of an individual are difficult to measure. The test–retest correlation for individual's olfactory threshold is often low (Punter, 1983). Even within an individual, threshold values will generally decrease with practice (Engen, 1960; Mojet et al., 2001; Rabin and Cain, 1986), and superimposed upon this practice effect is a high level of seemingly random variation (Stevens et al., 1988). Individuals may become sensitive to odorants to which they were formerly anosmic, apparently through simple exposure (Wysocki et al., 1989). Increased sensitivity as a function of exposure may be a common phenomenon among women of childbearing age (Dalton et al., 2002; Diamond et al., 2005).

Sensory adaptation and momentary changes in sensitivity due to sequences may have occurred in

the experiments of Stevens et al. (1988) and could have contributed to some instability in the measurements. As predicted by sequential sensitivity analysis (Masuoka et al., 1995; O'Mahony and Odbert, 1985) the specific stimulus sequence will render discrimination more or less difficult. After the stronger of two stimuli, an additional second strong stimulus presented next may be partially adapted and seem weaker than normal. Stevens et al. remarked that sometimes subjects would get all five pairs correct at one level with some certainty that they "got the scent" but lost the signal at the next level before getting it back. This report and the reversals of performance in threshold data are consistent with adaptation effects temporarily lessening sensitivity. The sensory impression will sometimes "fade in and out" at levels near threshold.

In attempts to avoid adaptation effects, other researchers have gone to fewer presentations of the target stimulus. For example, Lawless et al. (1995) used one target among three blank stimuli, a 4-AFC test that has appeared in previous studies (e.g., Engen, 1960; Punter, 1983). This lowers the chance performance level and lessens the potential adaptation at any one concentration step. To guard against the effects of correct guessing, threshold was taken as the lowest concentration step with a correct choice when all higher concentrations were also correct. Thresholds were measured in duplicate ascending runs in a test session, and a duplicate session of two more ascending runs was run on a second day. Correlations across the four ascending series ranged from 0.75 to 0.92 for cineole and from 0.51 to 0.92 for carvone. For carvone, thresholds were better duplicated within a day ($r = 0.91$ and 0.88) than across days (r from 0.51 to 0.70). This latter result suggests some drift over time in odor thresholds, in keeping with the variability seen by Stevens et al. (1988). However, results with this ascending method may not be this reliable for all compounds. Using the ascending 4-AFC test and a sophisticated olfactometer, Punter (1983) found median retest correlations for 11 compounds to be only 0.40. The sense of taste may fare somewhat better. In a study of electrogustometric thresholds with ascending paired tests requiring five correct responses, retest correlations for an elderly population were 0.95 (Murphy et al., 1995).

In many forced-choice studies, high variability in smell thresholds is also noted across the testing pool. Brown et al. (1978) stated that for any test compound, a number of insensitive individuals would likely be seen in the data set, when 25 or more persons were tested to determine an average threshold. Among any given pool of participants, a few people with otherwise normal smell acuity will have high thresholds. This is potentially important for sensory professionals who need to screen panelists for detection of specific flavor or odor notes such as defects or taints. In an extensive survey of thresholds for branched-chain fatty acids, Brennand et al. (1989) remarked that "some judges were unable to identify the correct samples in the pairs even in the highest concentrations provided" and that " panelists who were sensitive to most fatty acids found some acids difficulty to perceive" (p. 109). Wide variation in sensitivity was also observed to the common flavor compound, diacetyl, a buttery-smelling by-product of lactic bacteria fermentation (Lawless et al., 1994). Also, simple exposure to some chemicals can modify specific anosmia and increase sensitivity (Stevens and O'Connell, 1995).

6.9 Alternative Methods: Rated Difference, Adaptive Procedures, Scaling

6.9.1 *Rated Difference from Control*

Another practical procedure for estimating threshold has involved the use of ratings on degree-of-difference scales, where a sample containing the to-be-recognized stimulus is compared to some control or blank stimulus (Brown et al., 1978; Lundahl et al., 1986). Rated difference may use a line scale or a category scale, ranging from no difference or "exact same" to a large difference, as discussed in Chapter 4. In these procedures ratings for the sensory difference from the control sample will increase as the intensity of the target gets stronger. A point on the plot of ratings versus concentration is assigned as threshold. In some variations on this method, a blind control sample is also rated. This provides the opportunity to estimate a baseline or false alarm rate based on the ratings (often nonzero) of the control against itself. Identical samples will often get nonzero difference estimates due to the moment-to-moment variability in sensations.

In one application of this technique for taste and smell thresholds, a substance was added in various levels to estimate the threshold in a food or beverage. In each individual trial, three samples would be compared to the control sample with no added flavor—two adjacent concentration steps of the target compound and one blind control sample (Lundahl et al., 1986). Samples were rated on a simple 9-point scale, from zero (no difference) to eight (extreme difference). This provided a comparison of the control to itself and a cost-effective way of making three comparisons in one set of samples. Since sample concentrations within the three rated test samples were randomized, the procedure was not a true ascending series and was dubbed the "semi-ascending paired difference method."

How is the threshold defined in these procedures? One approach is to compare the difference ratings for a given level with the difference ratings given to the control sample. Then the threshold can be based on some measure of when these difference ratings diverge, such as when they become significantly different by a *t*-test (see Brown et al., 1978). Another approach is simply to subtract the difference score given to the blind control from the difference score give to each test sample and treat these adjusted scores as a new data set. In the original paper of Lundahl et al. (1986), this latter method was used. In the analysis, they performed a series of *t*-tests. Two values were taken to bracket the range of the threshold. The upper level was the first level yielding a significant *t*-test versus zero, and the lower level was the nearest lower concentration yielding a significant *t*-test versus the first. This provided an interval in which the threshold (as defined by this method) lies between the two bracketing concentrations.

One problem with this approach is that when the threshold is based on the statistical significance of *t*-statistics (or any such significance test), the value of threshold will depend upon the number of observations in the test. This creates a nonsensical situation where the threshold value will decrease as a function of the number of panelists used in the test. This is an irrelevant variable, a choice of the experimenter, and has nothing to do with the physiological sensitivity of the panelist or the biological potency of the substance being tested, a problem recognized by Brown et al. (1978) and later by Marin et al. (1991). Marin et al. also pointed out that a group threshold, based on a larger number of observations than an individual threshold, would be lower than the mean of the individual thresholds, due to the larger number of observations, another oddity of using statistical significance to determine the threshold.

Instead of using statistical significance as a criterion, Marin et al. determined the point of maximum curvature on the dose–response curve as the threshold.

Such an approach makes sense from consideration of the general form of the dose–response (psychophysical) curve for most tastes and odors. Figure 6.7 shows a semi-log plot for the Beidler taste equation, a widely applied dose–response relationship in studies of the chemical senses (see Chapter 2). This function has two sections of curvature (change in slope, i.e., acceleration) when plotted as a function of log concentration. There is a point at which the response appears to be slowly increasing out from the background noise and then rises steeply to enter the middle of the dynamic range of response. The point of maximum curvature can be estimated graphically or determined from curve fitting and finding the maximum rate of change (i.e., maximum of the second derivative) (Marin et al., 1991).

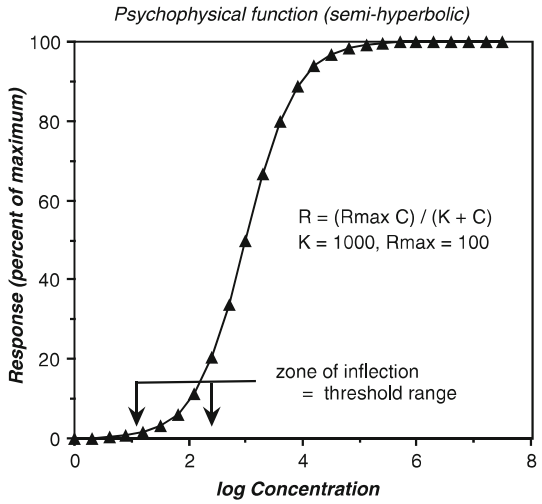

Fig. 6.7 Beidler curve.

6.9.2 Adaptive Procedures

Popular methods for threshold measurement for visual and auditory stimuli have been procedures in which the next stimulus intensity level to be tested depends

upon detection or non-detection at the previous interval. In these procedures, the subject will track around the threshold level, ascending in intensity when performance is incorrect (or non-detection is the response) and descending in physical intensity when performance is correct (or detection is indicated). A common example of this procedure is in some automated hearing tests, in which the person being tested pushes a button as long as the signal is not audible. When the button is depressed, the tone will increase in intensity and when the button is released, the tone will decrease in intensity. This automated tracking procedure leads to a series of up and down records, and an average of reversal points is usually taken to determine the threshold. Adaptive procedures may be more efficient than a traditional procedure like the method of limits. They focus on the critical range around threshold and do not waste time testing intensity levels very much higher or very much lower than the threshold (McBurney and Collings, 1977). Further information on these methods can be found in Harvey (1986).

With discrete stimuli, rather than those that are played constantly as in the example of the hearing test, the procedure can be used for the taste and smell modalities as well. This procedure is sometimes called a staircase method, since the record of ascending and descending trials can be connected on graph paper to produce a series of step intervals that visually resemble a staircase. An example is shown in Fig. 6.8. The procedure creates a dependence of each trial on previous trials that may lead to some expectations and bias on the part of the respondent. Psychophysical researchers have found ways to undo this sequential dependence to counteract observer expectancies. One example is the double random staircase procedure (Cornsweet, 1962) in which trials from two staircase sequences are randomly intermixed. One staircase starts above the threshold and descends, while the other starts below the threshold and ascends. On any given trial, the observer is unaware which of the two sequences the stimulus is chosen from. As in the simple staircase procedure, the level chosen for a trial depends upon detection or discrimination in the previous trial, but of that particular sequence. Further refinements of the procedure involve the introduction of forced-choice (Jesteadt, 1980) to eliminate response bias factors involved in simple yes/no detection.

Another modification to the adaptive methods has been to adjust the ascending and descending rules so that some number of correct or incorrect judgments is required before changing intensity levels, rather than the one trial as in the simple staircase (Jesteadt, 1980). An example is the "up down transformed response" rule or UDTR (Wetherill and Levitt, 1965). Wetherill and Levitt gave an example where two positive judgments were required before moving down, and only one negative judgment at a given level before moving up. An example is shown in Fig. 6.9. Rather than estimating the 50% point on a traditional psychometric function, this more stringent requirement now tends to converge on the 71% mark as an average of the peaks and valleys in the series. A forced choice can be added to an adaptive procedure. Sometimes

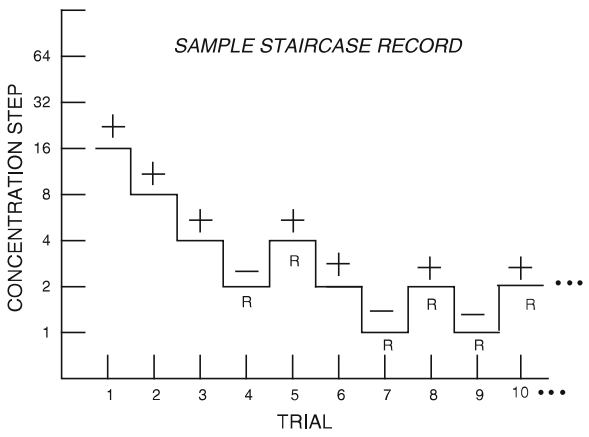

Fig. 6.8 Staircase example.

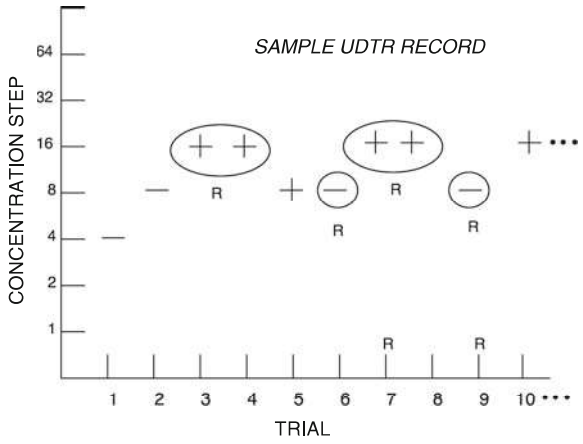

Fig. 6.9 Staircase example.

the initial part of the test sequence is discarded in analysis as it is unrepresentative of the final threshold and derives from a time when the observer may still be warming up to the test procedure. Examples of the up-down procedure can be found in the literature on PTC/PROP tasting (e.g., Reed et al., 1995). Recent advances in adaptive methods have shown that thresholds may be estimated in very few trials using these procedures, a potential advantage for taste and smell measurement (Harvey, 1986; Linschoten et al., 1996).

6.9.3 Scaling as an Alternative Measure of Sensitivity

Threshold measurements are not the only way to screen individuals for insensitivity to specific compounds like PTC or to screen for specific anosmia. Do thresholds bear any relation to suprathreshold responding? While it has been widely held that there is no necessary relationship between threshold sensitivity and suprathreshold responding (Frijters, 1978; Pangborn, 1981), this assertion somewhat overstates the case. Counter-examples of good correlations can be seen in tests involving compounds like PTC where there are insensitive groups. For example, there is a –0.8 correlation between simple category taste intensity ratings for PTC and the threshold, when the rated concentration is near the antimode or center between the modes of a bimodal threshold frequency distribution (Lawless, 1980). Thus ratings of a carefully chosen level can be used for a rapid screening method for PTC taster status (e.g., Mela, 1989).

Similar results have been noted for smell. Berglund and Högman (1992) reported better reliability of suprathreshold ratings than threshold determinations in screening for olfactory sensitivity. Stevens and O'Connell (1991) used category ratings of perceived intensity as well as qualitative descriptors as a screening tool before threshold testing for specific anosmia. Threshold versus rating correlations were in the range of –0.6 for cineole, –0.3 for carvone, and –0.5 for diacetyl (Lawless et al., 1994, 1995). The correlations were obtained after subtraction of ratings to a blank stimulus, in order to correct for differences in scale usage. Thus there is a moderate negative correlation of sensitivity and rated intensity when one examines the data across a highly variable group as is the case with specific anosmia or tasting PTC bitterness. The correlation is negative since higher thresholds indicate lower sensitivity and thus lower rated intensity.

6.10 Dilution to Threshold Measures

6.10.1 Odor Units and Gas-Chromatography Olfactometry (GCO)

In this section, several applied methods will be described that make use of the threshold concept in trying to determine the sensory impact of various flavors and odor materials. The first group of methods concerns the olfactory potency of volatile aroma compounds as they are found in foods or food components. The issue here becomes one of not just threshold determination, but determination of both the threshold and the actual concentration present in a food sample. The ratio of these concentrations (actual concentration to threshold concentration) can help indicate whether or not a given flavor substance is likely to contribute to the overall sensory impression in a food. These ratios are commonly called "odor units." The second much older method is similar in logic and was developed to determine the point at which the irritative or heat sensations from pepper compounds would be first detectable when diluted to a given extent, the Scoville procedure. Both of these techniques then use dilution-to-threshold as a measure of sensory impact.

When a complex natural product like a fruit extract is analyzed for its chemical components, hundreds or even thousands of chemicals may be identified, many of which have odor properties. The number of potential flavor compounds identified in any product seems only to be limited by the resolution and sensitivity of the current methods in analytical chemistry. These methods are always improving leading to longer and longer lists of possible contributing flavor materials (Piggott, 1990). Flavor scientists need to find a way to narrow the list or to separate those compounds which are most likely contributing to the overall flavor from those compounds that are present in

such low concentrations that they are probably not important. Obviously, a sensory-based method is needed in conjunction with the analytical chemistry to provide a bioassay for possible sensory impact (Acree, 1993).

Thresholds can be useful in addressing this kind of problem. The reasoning goes that only those compounds that are present in the product in concentrations above their threshold are likely to be contributors to the flavor of the product. There are a number of potential flaws in this thinking discussed below, but for now let us see how this can be put to use. Given a concentration C present in a natural product, a dimensionless quantity can be derived by dividing that concentration by the threshold concentration C_t, and the ratio C/C_t defines the number of odor units (or flavor units) for compounds assessed by smell. According to this logic, only those compounds with odor units greater than one will contribute to the aroma of the product. This reasoning is extended sometimes to include the idea that the greater the number of odor units, the greater the potential contribution. However, it is now widely recognized that the odor unit is a concentration multiple and not a measure of subjective magnitude. Only direct scaling methods can assess the actual magnitude of sensation above threshold and the psychophysical relationship between concentration and odor intensity (Frijters, 1978). Furthermore, this idea ignores the possibility of subthreshold additivity or synergy (Day et al., 1963). A closely related group of chemical compounds might all be present below their individual thresholds, but together could stimulate common receptors so as to produce an above-threshold sensation. Such additivity is not predicted by the odor unit approach and such a group of compounds could be missed in dilution analysis.

Nonetheless, thresholds provide at least one isointense reference point on the dose response curve, so they have some utility as a measure of potency used to compare different odor compounds. In analyzing a food, one could look up literature values for all the identified compounds in the product in one of the published compendia of thresholds (e.g., ASTM, 1978; van Gemert, 2003). If the concentration in the product is determined, then the odor unit value can be calculated by simply dividing by threshold. However, it is important to remember that the literature values for thresholds depend upon the method and the medium of testing. Unless the same techniques are used and the same medium was used as the carrier (rarely the case) the values may not be necessarily comparable for different compounds.

A second approach is to actually measure the dilutions necessary to reach threshold for each compound, starting with the product itself. This necessitates the use of a separatory procedure, so that each compound may be individually perceived. The combination of gas chromatography with odor port sniffing of a dilution series is a popular technique (Acree, 1993). Various catchy names have been applied to such techniques in the flavor literature, including Aroma Extract Dilution Analysis (for examples, see Guth and Grosch, 1994; Milo and Grosch, 1993; Schieberle and Grosch, 1988), CHARM analysis (Acree et al., 1984) or more generically, gas chromatography olfactometry or GCO. The basis of these techniques is to have subjects respond when an odor is perceived when sniffing the exit port during a GC run. In recent years, the effluent has been embedded in a cooled, humidified air stream to improve the comfort of the observer and to increase sensory resolution of the eluting compounds. Over several dilutions, the response will eventually drop out, and the index of smell potency is related to the reciprocal of the dilution factor. The sniffer's responses occur on a time base that can be cross-referenced to a retention index and then the identity of the compound can be determined by a combination of retention index, mass spectrometry, and aroma character. In practice, these techniques considerably shorten the list of potential aroma compounds contributing to flavor in a natural product (e.g., Cunningham et al., 1986).

The method has also been used as an assessment technique for measuring the sensitivity of human panelists, as opposed to being a tool to determine the sensory impact of flavor compounds (Marin et al., 1988). In this approach, the gas chromatograph once again serves as an olfactometer. Known compounds can be presented as a dilution series of mixtures. Variation in individual thresholds can readily be assessed for a variety of compounds since they can be combined in a GC run as long as they have different retention times, i.e., do not co-elute on the column of choice. The potential use of GCO for screening panelists, for assessing odor responses in general, and for assessing specific anosmias has also been attempted (Friedrich and Acree, 2000; Kittel and Acree, 2008).

6.10.2 Scoville Units

Another example of a dilution method is the traditional Scoville procedure for scoring pepper heat in the spice trade. This procedure was named for W. Scoville who worked in the pharmaceutical industry in the early twentieth century. He was interested in the topical application of spice compounds like pepper extracts as counterirritants, and he needed to establish units that could be used to measure their potency. His procedure consisted of finding the number of dilutions necessary for sensations to disappear and then using this number of dilutions as an estimate of potency. In other words, potency was defined as a reciprocal threshold. A variation of this procedure was adopted by the Essential Oil Association, British Standards Institution, International Standards Organization, American Spice Trade Association (ASTA), and adopted as an Indian standard method (for a review, see Govindarajan, 1987).

The procedure defined units of pungency as the highest dilution at which a definite "bite" would be perceived and thus contains instructions consistent with a recognition threshold. Scoville units were dilution factors, now commonly given as mL/g. ASTA Method 21 (ASTA, 1968) is widely used and contains some modifications in an attempt to overcome problems with the original Scoville procedure. In brief, the method proceeds as follows: Panelists are screened for acuity relative to experienced persons. Dilution schedules are provided which simplify calculations of the eventual potency. Solutions are tested in 5% sucrose and negligible amounts of alcohol. Five panelists participate and concentrations are given in ascending order around the estimated threshold. Threshold is defined as the concentration at which three out of five judges respond positively.

This method is difficult in practice and a number of additional variations have been tried to improve on the accuracy and precision of the method (Govindarajan, 1987). Examples include the following: (1) substitution of other rules for 3/5, e.g., mean + SD of 20–30 judgments, (2) use of a triangle test, rather than simple yes/no at each concentration, (3) requiring recognition of pungency (Todd et al., 1977), (4) reduction of sucrose concentration in the carrier solution to 3%, and (5) use of a rating scale from 1 (definitely not detectable) to 6 (definitely detectable). This latter modification defined a detection threshold at mean scale value of 3.5. Almost all these methods specify mandatory rest periods between samples due to the long lasting nature of these sensations. The measurements are still difficult. One problem is that capsaicin, the active heat principle in red pepper, is prone to desensitize observers within a session and also regular consumers of hot spices also become less sensitive, leading to wide individual differences in sensitivity among panelists (Green, 1989; Lawless et al., 1985). Alternative procedures have been developed based on rating scales with fixed physical references (Gillette et al., 1984) and these have been endorsed by ASTM (2008c) as standard test methods. These rating scale procedures show good correlations with instrumental measures of capsaicin content in pepper samples and can be cross-referenced to Scoville units for those who prefer to do business in the traditional units (Gillette et al., 1984).

6.11 Conclusions

Threshold measurements find three common uses in sensory analysis and flavor research. First, they can be used to compare the sensitivities of different panelists. Second, they can be used as an index of the biological potency of a flavor compound. Third, they can provide useful information regarding the maximum tolerable levels of an off-flavor or taint. A variety of different techniques have been used to find thresholds or have employed the threshold concept in practical flavor work. Examples of different threshold methods are given in Table 6.4. In spite of their practical applications, the usefulness of threshold measures is often questioned in sensory evaluation. One criticism is that thresholds are only one point on an intensity function and thus they do not tell us anything about above-threshold responding. There are some good examples in which thresholds do not predict or do not correlate very well with suprathreshold responses. For example, patients irradiated for cancer may lose their sense of taste temporarily and thresholds return to normal long before suprathreshold responsiveness is recovered (Bartoshuk, 1987). However, as we have seen both in the case of PTC tasting and in specific anosmias, insensitive individuals (as determined by their threshold) will also tend to be less responsive above