INTERNATIONAL STANDARD

ISO 13301

First edition
2002-09-15

Sensory analysis — Methodology — General guidance for measuring odour, flavour and taste detection thresholds by a three-alternative forced-choice (3-AFC) procedure

Analyse sensorielle — Méthodologie — Lignes directrices générales pour la mesure des seuils de détection d'odeur, de flaveur et de goût par une technique à choix forcé de 1 parmi 3 (3-AFC)

Reference number
ISO 13301:2002(E)

© ISO 2002

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

© ISO 2002

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.ch
Web www.iso.ch

Printed in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 3.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this International Standard may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 13301 was prepared by Technical Committee ISO/TC 34, *Food products*, Subcommittee SC 12, *Sensory analysis*.

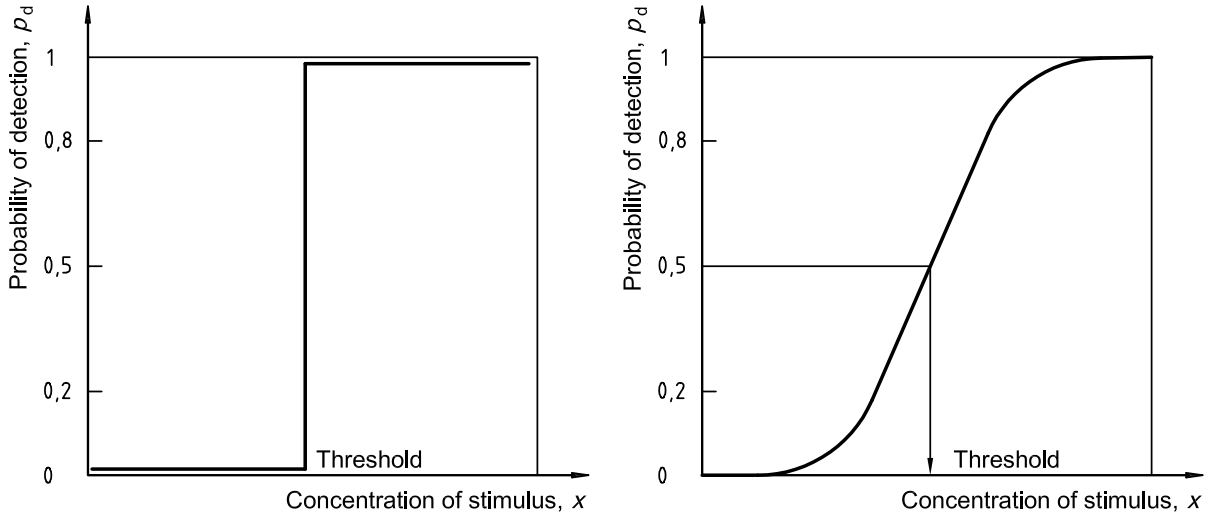
Annexes A and B of this International Standard are for information only.

© ISO 2002 – All rights reserved

Introduction

The concept of “threshold” has wide use in sensory analysis and is often used in the literature on sensory studies of food and drink. Data on sensory thresholds to chemical stimuli are used in sensory studies in two main ways: as measures of the sensitivity of assessors or groups of assessors to specific stimuli; as measures of the ability of chemical substances to evoke sensory responses in assessors. In the first, the value of the threshold is taken as a description of an assessor’s performance; in the latter, as a measure of a property of the substance.

The term “threshold” was introduced by 19th century psychophysicists and used to denote a stimulus concentration above which the stimulus could be detected, and below which it could not (see Figure 1a).



a) Traditional notion of threshold **b) Probabilistic nature of threshold**

Figure 1 — Traditional notion and probabilistic nature of threshold

However, in practice the graph of the probability of detection[1)] against the intensity of the stimulus is always an ogive (see Figure 1b), and it is convenient to assume, for purposes of calculation, that the threshold fluctuates so that a particular stimulus concentration exceeds it on some occasions but not on others. The threshold can then be obtained as an estimate of the median of these momentary values, i.e. as the stimulus concentration for which the probability of detection is 0,5. The threshold defined in this way has analogies with median effect levels used in branches of biology such as pharmacology and toxicology, which are concerned with the effect of chemicals on organisms.

1) This International Standard is based on the use of the 3-AFC method of presenting the stimuli, and the probability of detection, p_d, is modeled as $p_d = 1{,}5 \times p_c - 0{,}5$, where p_c is the probability of a correct selection. This is strictly a “guessing model” of the assessor’s behaviour. It is not a psychometric model of the assessor’s decision process, such as a Signal Detection model, which could also be applied, see Macmillan and Creelman [13].

Where detection thresholds of a particular substance in air or water have been measured in more than one laboratory, the reported values often span two or three orders of magnitude or more (Devos *et al.* [6], Fazzalari [10], van Gemert *et al.* [14]). This range is greater than can be expected from experimental errors alone or from differences in the processing of data; but it probably can be accounted for by difference in concepts of thresholds between laboratories, and differences in experimental procedure. Devos *et al.* [6] suggest a procedure for standardizing detection thresholds in air.

The user needs to be aware that the determination of detection thresholds requires more experimental effort than is at first apparent from this description. Experimental results demonstrate that on repeated testing, the observed individual thresholds tend to decrease, and the difference between individuals likewise tends to decrease. Threshold testing is often an unfamiliar activity, and assessors will improve their sensitivity as they become accustomed to the substance and the mechanics of the test. The 3-AFC procedure requires that assessors can recognize the stimulus. Training programmes require effort but will in turn yield needed information about each assessor's range of partial detection. Results improve as the experimenter learns to tailor the concentrations presented to each assessor's range, see 6.3.

© ISO 2002 – All rights reserved

Sensory analysis — Methodology — General guidance for measuring odour, flavour and taste detection thresholds by a three-alternative forced-choice (3-AFC) procedure

1 Scope

This International Standard provides guidance on:

— obtaining data on the detection of chemical stimuli that evoke responses to odour, flavour and taste by a 3-AFC (three-alternative forced-choice) procedure;

— the processing of the data to estimate the value of a threshold and its error bounds, and other statistics related to the detection of the stimulus.

Typically, the procedures will be used in one of the following two modes:

— investigation of the sensitivity of assessors to specific stimuli;

— investigation of the ability of a chemical substance to stimulate the chemoreceptive senses.

(Although experiments may encompass both modes.)

Examples of the first mode would include studies of the differences among individuals or specified populations of individuals in sensitivities and of the effects of age, gender, physiological condition, disease, administration of drugs and ambient conditions on sensitivity. Examples of the latter mode would include:

— studies in flavour chemistry and the impact of specified chemicals on the flavour of foods;

— classification of chemicals for their impact on humans, if present in the environment;

— studies on the relationship of molecular structure to capacity of a chemical to act as a stimulant;

— quality assurance of gaseous effluents and of water, foods and beverages;

— studies in the mechanism of olfaction.

In both modes the way in which probability of a correct response changes with intensity of stimulus, i.e. the slope of the dose/response curve, could be an important aspect of the study as well as the threshold value, and the data processing procedures described here provide this information.

The focus of this International Standard is on data requirements and on computational procedures. Regarding the validity of the data, the text is restricted to general rules and precautions. It does not differentiate between detection and difference thresholds; fundamentally, the procedures measure a difference threshold because a test sample is compared with a reference sample. Typically, the reference sample is not intended to contain the stimulus under investigation, but the Guidelines do not exclude experimental design in which the reference could contain the stimulus, or it might not be known if the reference contains the stimulus. The Guidelines do not measure a recognition threshold as defined in ISO 5492. They do not address the standardization of methods of determining air quality as a European Standard is in preparation [9].

2 Normative references

The following normative documents contain provisions which, through reference in this text, constitute provisions of this International Standard. For dated references, subsequent amendments to, or revisions of, any of these publications do not apply. However, parties to agreements based on this International Standard are encouraged to investigate the possibility of applying the most recent editions of the normative documents indicated below. For undated references, the latest edition of the normative document referred to applies. Members of ISO and IEC maintain registers of currently valid International Standards.

ISO 5492:1992, *Sensory analysis — Vocabulary*

ISO 6658:1985, *Sensory analysis — Methodology — General guidance*

ISO 8586-1:1993, *Sensory analysis — General guidance for the selection, training and monitoring of assessors — Part 1: Selected assessors*

ISO 8586-2:1994, *Sensory analysis — General guidance for the selection, training and monitoring of assessors — Part 2: Experts*

ISO 8589:1988, *Sensory analysis — General guidance for the design of test rooms*

3 Terms and definitions

For the purposes of this International Standard, the terms and definitions given in ISO 5492, as well as the following, apply.

3.1
stimulus
substance that may or may not cause a sensation, detectable by one or more of the senses, depending on the amount present

3.2
medium
any material used to dissolve, dilute, disperse or sorb a stimulus whose threshold is to be measured

3.3
reference sample
quantity of the medium containing no added stimulus

3.4
test sample
quantity of the medium to which a stimulus has been added at a known concentration

3.5
three-alternative forced-choice (3-AFC) test
test of discrimination in which the assessor is presented with three samples, one of which is a test sample containing a nominated stimulus familiar to the assessor, the other two being references, and where the assessor is instructed to indicate the test sample

3.6
presentation
set of three samples forming a 3-AFC test

3.7
threshold model
model of sensory detection where a stimulus presented on a particular trial is either detected (resulting in a correct response) or is not detected (resulting in a response being made at random)

3.8
signal-detection model
model of sensory detection where a stimulus presented on a particular trial provides some level of evidence of its presence

NOTE The evidence contributes to a decision by the assessor about the presence or absence of the stimulus.

© ISO 2002 – All rights reserved

3.9
detection threshold
the lowest intensity of a sensory stimulus that has a probability of detection of 0,5 under the conditions of the test, as calculated from the threshold model

3.10
individual threshold
detection threshold of a single assessor

3.11
average threshold
average (whose type must be specified, e.g. arithmetic mean, geometric mean, or median) of individual thresholds

3.12
group threshold from pooled data
estimate obtained by using the sum of outcomes for a particular group of assessors at each concentration of the stimulus as input when fitting the statistical model

4 Principles

4.1 Experimental procedures

The stimulus is formulated in the medium at a specified concentration and is presented along with a pair of reference samples to the assessor. The assessor is required to select one of the samples as containing the stimulus or having the stimulus at a greater concentration. The assessor must make a selection. It is a requirement of the 3-AFC test that the assessor be able to recognize the stimulus.

Typically the stimulus is dissolved in air or water. It is unlikely that a gas other than air will be used as a gaseous medium in tests with human assessors, but solvents other than water, solutions in water or other solvents, or solids, e.g. foods, can be used as liquid or solid medium to dilute the stimulus as the experiment dictates. It is essential that the medium be homogeneous so that the members of the pair of references are identical, and the same in all presentations.

The stimulus is presented at several concentrations. The presentations are replicated, at each concentration, a sufficient number of times to achieve a desired precision of the threshold and parameters of the mathematical model. The nature of the replications within assessors, across assessors, and combinations of the two are set by the experimental design of the study.

4.2 Data processing

The outcome of a presentation is a binary result – the sample nominated by the assessor is the test sample (a correct selection) or is one of the references (an incorrect selection). The number of correct selections is summed over the number of presentations at each stimulus concentration and forms, along with the total number of presentations and the stimulus concentration, the data to be processed for obtaining the derived statistics. The statistical model is that the number of correct selections at a particular concentration comes from a binomial distribution.

For the 3-AFC test, the threshold is the concentration of the stimulus at which the proportion of correct selections is equal to 2/3, i.e. 50 % above chance. The data, as proportions of correct selections, can simply be inspected and interpolated to derive this point, but a more accurate estimate of the threshold, and its bounds, can be obtained by fitting a mathematical model to the data. A logistic model is used in these guidelines, and the model is fitted by a maximum likelihood procedure, or alternatively, by a least squares procedure. The fitting estimates the two parameters of the model, one a location parameter, the other a shape parameter. The former locates the fitted curve on the stimulus continuum, the latter determines the steepness of the curve. The fitted curve allows estimates of proportions of detection other than 50 % to be derived.

The simplest model to fit is one in which the distribution of proportion of correct selections comes from a single, approximately normal, distribution. This would typically be the case where the data come from replications within a single assessor. A single logistic function can then be adequately fitted, that is, one with a single pair of values for

the parameters of the curve. It is not uncommon for the sensitivities to chemicals to be not normally distributed, or even symmetrically distributed, among assessors. For some chemical stimuli the distributions are distinctly bimodal, but deviations from a normal distribution are difficult to demonstrate unless measurements are made with a large sample of assessors, typically more than 100. A single logistic function will not be an adequate fit to data that come from a distribution which deviates significantly from a single, normal distribution, but the mathematical model can be extended to accommodate these cases.

5 Experimental procedures

5.1 Preparation of samples

5.1.1 General precautions

See ISO 6658. Ascertain that stimulus and medium are stable over the duration of the study and are non-toxic and nonallergenic. Ascertain that they are representative of the purpose of the study, e.g. exhaust gases may vary with the process generating them, and chemical substances may require purification to remove off-flavours or irritants from the molecule to be studied. Prepare a large enough homogeneous quantity of both stimulus and medium to ensure that assessors receive identical presentations with exception of the concentration of stimulus and its position in the set. Prepare the samples in a facility that complies with ISO 8589. Use containers that do not adsorb the test chemical or contribute odour or taste. Make certain that the presence or absence of the stimulus cannot be detected visually or by any means available to an assessor other than the chemical senses. Store samples away from light and heat when not in use.

5.1.2 Gases

Collect or prepare stimulus and medium in vessels such as teflon- (PTFE) coated bottles or balloons. If the stimulus is an inodorous gas containing an odorous impurity, flush the vessel and associated tubing and valves several times with a fresh sample in order to saturate the walls. For the same reason, and to avoid volume changes, maintain a constant temperature near that to be used when presenting the gases to the assessors. Use smoothbore PTFE-coated tubing and valves free from points of sudden pressure change.

5.1.3 Liquids

For stimuli to be presented in an aqueous medium, make certain that complete dissolution can be obtained and maintained for the duration of the experiment. For partially hydrophobic substances, prepare the first dilution stage in ethanol or ethylene glycol purified with activated carbon to remove off-odours. Note that distilled water and absolute alcohol often contain strong odours; use food grade product instead and purify with activated carbon if required. Present fully hydrophobic substances in a nonaqueous solvent such as odourless liquid paraffin or dinonyl phthalate and avoid plastic containers as the substance may dissolve in the polymer. When preparing sequential dilutions, be aware that the higher the dilution, the larger the proportion of the stimulus that may be lost by adsorption to the vessel wall. As far as is possible, prepare each dilution by microsyringe or equivalent, directly from a stock solution, and avoid sequences of preparing each dilution from the preceding sample.

5.1.4 Solids

The medium of interest is typically a food such as cheese, fish or meat. Unless a technique exists whereby the solid can be dissolved and reconstituted, finely divide or comminute it before adding the stimulus in a suitable solvent, then mix well and allow time for the chemical to diffuse within the matrix before preparing the samples for presentation to the assessors. Code each aliquot, e.g. with a random, 3-digit number.

5.2 Selection of concentrations of the stimulus

Present a series of 3-AFC presentations of which each concentration is greater than the preceding one by approximately a factor denoted by X. Be guided by the acceptable size of the error of the threshold estimate: typically choose $X \approx$ 3-5 for approximate studies and $X \approx 2$ for higher precision. For each assessor, choose a strategy of experimentation that will result in defining the ogive of the logistic model at points distributed over his or her range of partial detection. The most effective data points are those corresponding to 45 % to 90 % correct selection in the test, i.e. p_d = 0,18 to 0,85.

© ISO 2002 – All rights reserved

For economy of sample and assessor's time, begin by locating the concentration range of interest for each assessor using a large factor X. Observe that these initial tests also serve to demonstrate the mechanics of the test and to teach the assessors how to recognize the stimulus when it is above their range of partial detection.

Proceed with the definitive set of 3-AFC presentations at concentrations tailored to each assessor using a low factor X. If on completion it is found that the data do not adequately define an assessor's ogive, administer additional concentration levels until this is the case. Regularly ask an assessor to describe the nature of the detected stimulus so as to guard against lapses of memory for it. Interrogation may also uncover an unintended sequence of correct replies caused by chance and not by detection; e.g. a series of 3 chance hits will occur once in 27 tests.

5.3 Presentation of samples

5.3.1 Preparation

Present samples with assessors seated in booths (see ISO 8589) and observe the rules of good sensory practice as described in ISO 6658. Code samples with three-digit random numbers, or place samples in a prearranged pattern, e.g. side-by-side in front of the assessor with the first sample on the left, using the identical pattern on the response sheet. To avoid positional bias, balance the three combinations of orders of presentation, AAB, ABA, BAA, across the assessors. Instruct assessors to minimize sensory fatigue by ingesting a minimum quantity of any sample that exhibits above-threshold concentration and by allowing sufficient time for sensory recovery between samples.

5.3.2 Gases

Present samples using an olfactometer such as those described in [8] and [12].

5.3.3 Liquids

Present non-volatile chemicals dissolved in purified water or in a flavourless solvent. Use containers that do not absorb the chemical, e.g. 100 ml glass beakers one quarter full. Present volatile chemicals in stoppered, wide mouthed containers suitable for sniffing or sipping, or in flexible closed containers, e.g. 250 ml squeeze bottles suitable for delivering a measured volume of headspace or liquid into the nostrils or mouth, see [4], [7] and [11]. If the medium is a beverage, use the type of container that is customary for sensory evaluation of the product.

5.3.4 Solids

If the medium is a food, present the samples in the form that is customary for sensory evaluation of the product.

5.4 Training of assessors

For most purposes, the threshold of interest is that of an informed observer, trained by repeated exposure to detect the substance in question whenever its presence is perceivable, e.g. as a pollutant in air or water, or as a component or taint of the flavour of a food or beverage. Familiarity with the substance is also a requirement in the 3-AFC test. Inadequate training may artificially extend the observed range of thresholds upwards by 1-2 orders of magnitude. An artificial extension downwards can result from overtraining, when assessors become adept at discovering the treated sample by means other than its flavour. If the threshold sought is that of a casual observer, e.g. for a warning agent in household gas, untrained assessors and mild distraction (e.g. noise) may be used and the triangle test or paired comparison substituted for the 3-AFC test.

A training programme can be by presentation of the stimulus monadically at high concentrations, then at two or more concentrations with the assessor requiring to rank them, then as 3-AFCs while locating the assessor's range of partial detection. Observe that initial thresholds decrease with practice and should tend to stabilize after 3 to 5 tests and that individual assessors may differ in their basic sensitivity to the substance in question by a factor of two or three orders of magnitude, or more.

5.5 Selection of assessors

5.5.1 General

Select assessors to meet the objectives of the investigation, following the guidelines given in ISO 8586-1 and ISO 8586-2.

5.5.2 Individual threshold

The test may be made, e.g. to compare an individual's threshold with a literature value, with a previously determined value under different circumstances, or with his or her thresholds for other substances. The test may be made to diagnose anosmia or hyperosmia, ageusia or hypergeusia.

5.5.3 Distribution of thresholds

The experimenter may wish to know the distribution of thresholds within a population. The group tested might itself be a sample drawn from a larger population, or it may be all members of a selected population, e.g. members of a testing panel. Selection of populations is outside the scope of this International Standard, but the experimenter should carefully define the population, or the sample of the population, under study. For the presentation of the results, see 6.7.

5.5.4 Measurement of thresholds of stimuli

The value of a group or average threshold for a stimulus is valid only for the panel of assessors used in the trials and the experimenter should be cautious in extrapolating the results outside of this panel. The experimenter should select the panel to meet the objectives and purposes of the measurements. For example, a study of the relative organoleptic properties of members of a set of chemicals could be carried out using a small panel of selected assessors, whereas a study of the properties of potential flavouring compounds in foods might require a larger panel which is representative of a particular population.

The number of assessors and the number of presentations to achieve a required precision of estimates are matters to be considered together. When small numbers of assessors are being used, it will be necessary to replicate presentations over assessors to generate sufficient data, whereas single presentations at each, or perhaps just some, concentrations to each assessor might be adequate for large panels.

5.6 Design of the experiment

5.6.1 Individual threshold

The most effective range of concentrations for estimating the parameters of the logistic is between 45 % and 90 % correct selections. Within this range the main determinant of precision of the estimates is the total number of presentations assuming they are roughly balanced around the threshold. Table 1 shows factors for approximate error bounds relative to the estimate of the threshold, in original concentration units. See also annex A.

Table 1 — Guide for determining the number of presentations required for a desired precision of an estimate of the threshold

Total number of presentations	40	60	80	100	120	160	200
Error bound relative to threshold	2,5	2,2	2,0	1,8	1,7	1,6	1,5

The bounds are obtained by both dividing and multiplying the estimate of the threshold by the factors in Table 1; e.g. if the threshold obtained with 80 presentations was 2,4 ppm (2,4ml/m^3), the bounds would be 1,2 ppm to 4,8 ppm. Precision increases only slowly above 200 presentations and the improvement is probably not worth the extra effort. A sequential strategy is effective. After a few replicate presentations at each concentration, fit the logistic and calculate the threshold and error bounds. Carry out more replicates at concentrations within the most effective range determined from the fitted logistic, and repeat until the desired precision is obtained.

© ISO 2002 – All rights reserved

5.6.2 Distribution of thresholds

Replicate the measurements in 5.6.1 over the selected assessors. Display the results in a histogram or in a cumulative frequency graph. Report the average threshold as the arithmetic mean, geometric mean or median, or if the distribution appears to be bimodal or multimodal, attempt to resolve the number of modes. For data processing, see clause 6.

5.6.3 Measurement of the threshold of a stimulus for a group of assessors

5.6.3.1 General

In choosing an experimental design, observe that variations in sensitivity between assessors is likely to be several fold greater than within an assessor. It follows that practical applicability of the resulting central value for the group is likely to be greater if replication is aimed at enlarging the number of assessors included in the test, rather than at increasing the number of presentations per assessor.

5.6.3.2 Group threshold from pooled data

Rather than separately fitting a logistic model to the data of each assessor, fit only a single logistic model to the pooled data, using all of the data at each given concentration as inputs into the model. Observe that the larger number of data obtained by pooling allows a better fit to be obtained for the pooled-group threshold, which is the detection threshold defined in 3.9. See the discussion in 6.5 and Examples B.2 and B.4 in annex B. Use this technique when differences between individuals are not a part of the experimental design, e.g. in classifying chemicals according to their importance as pollutants or sensory taints.

5.6.3.3 Average threshold

Replicate the measurements in 5.6.1 over the selected assessors. Display the results in a histogram as shown in Figure 2, or in a cumulative frequency graph. Use this technique when differences between individuals form a part of the objective of the study, e.g. in studying the impact of a flavour compound, a pollutant or a sensory taint on a particular population.

In Figure 2, the upper two histograms show the same 443 assessors. The bottom histogram has only 222 assessors, hence the vertical scale is doubled for comparability. Dilution step “0” represents the saturated solution for each odourant and hence the highest threshold (from Amoore [3]).

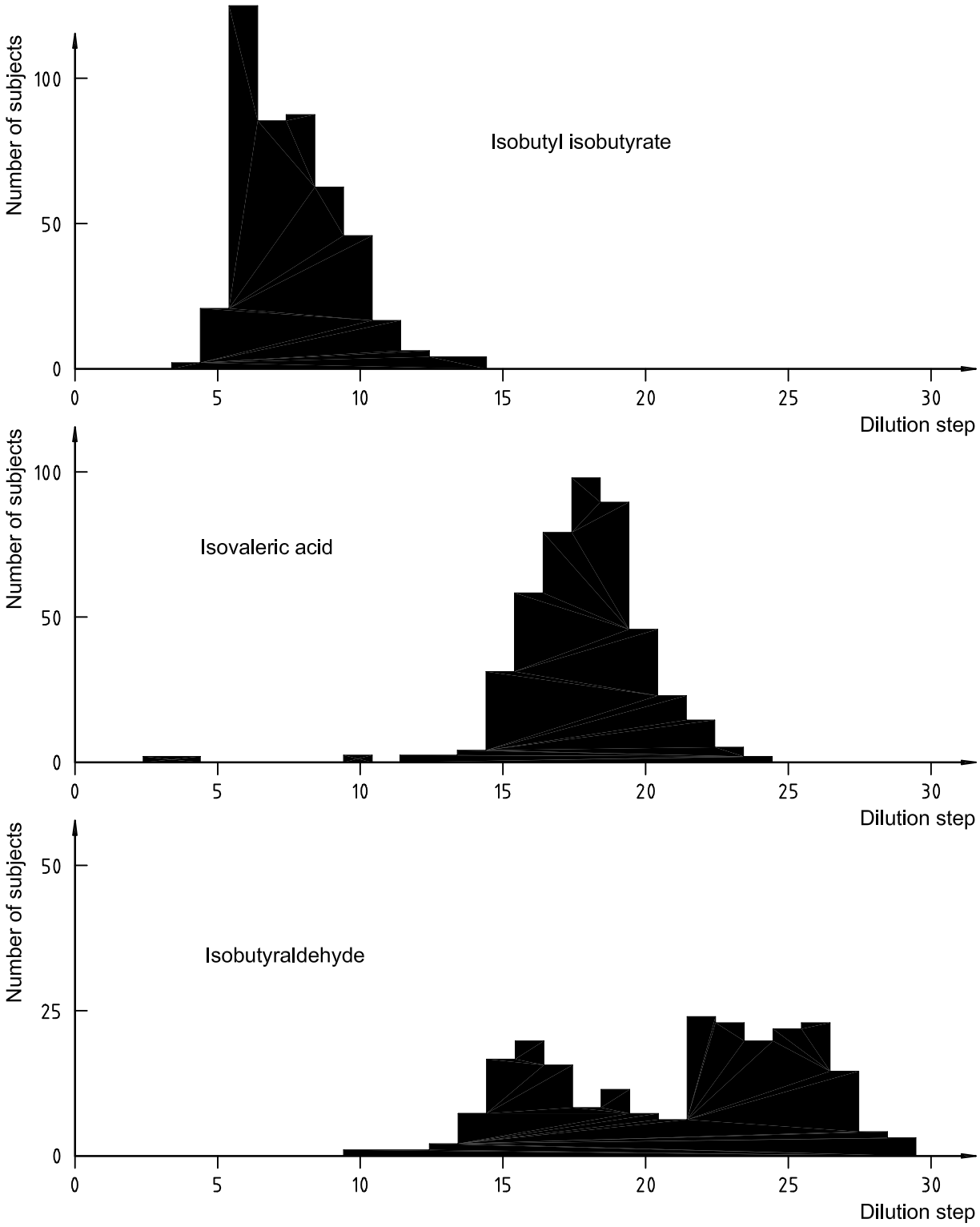


Figure 2 — Olfactory threshold distributions in the population

© ISO 2002 – All rights reserved

6 Data processing

6.1 The mathematical and statistical models

In a 3-AFC task, the probability of a randomly-selected response being correct is 1/3, or approximately 0,33, as only one of three available choices is correct. According to the threshold model, the probability of a correct response, p_c, is therefore related to the probability of detection, p_d, by equation (1)

$$p_c = p_d + \frac{1}{3}(1 - p_d) = \frac{2}{3} p_d + \frac{1}{3} \quad (1)$$

because on a proportion, p_d, of trials the stimulus is detected and a correct response is given while on the remaining proportion, $1 - p_d$, the assessors cannot detect the stimulus and must guess at random with a 1/3 probability of giving a correct answer. Because the threshold is defined as the stimulus concentration for which $p_d = 0,5$, it follows that it is the value for which $p_c = 0,67$.

The quantity p_c is observed data, whereas p_d is an inference from the threshold model. Of interest here is the inverse calculation:

$$p_d = 1,5 p_c - 0,5 \quad (2)$$

When the proportion of correct choices, p_c, in a series of difference tests repeated several times at each of several stimulus concentrations, is plotted against the concentrations, the points approximate to an ogive. If p_c is converted to p_d by equation (2), the graph of p_d forms another ogive (see Figure 1b) with asymptotes at 0 and 1 for sufficiently low and high stimulus concentrations. In the case of sensitivity to chemicals, intensity is usually expressed as the logarithm of concentration or dilution.

The ogive relating p_d to concentration can be modelled by the cumulative normal distribution or, more conveniently, by the cumulative logistic distribution, whose equation can be written:

$$p_c = \frac{\frac{2}{3}}{1 + e^{b(t-x)}} + \frac{1}{3} \quad (3)$$

with the stimulus concentration denoted by x while the values of the coefficients t and b depend on the data. When $x = t$, $p_d = 0,5$, so t is the threshold value of the stimulus. The parameter b determines the size of change in x required to produce any particular change in p_d and so determines the steepness of the ogive.

6.2 Preliminary inspection of data

6.2.1 Preparation

Make a preliminary inspection of the data, numerically or by a graph of proportion of correct responses, p_c, against log concentration. Note whether the results appear to conform to an ogive and whether the concentrations tested lie both above and below the estimated threshold, as they should for the estimate to be accurate. Obtain more data if this is not the case. Estimate the threshold visually and decide if this is accurate enough for the purpose for which it is required. If this is not the case, proceed to fit a model by hand calculator using the logit transformation described in 6.2.2, or use the maximum likelihood procedure described in 6.3 and in more detail in Examples B.2 to B.4.

6.2.2 Preliminary estimation of threshold and slope using the logit transformation

The logit is the equivalent of the familiar probit except that the latter is based on the cumulative normal distribution. Transform p_d to its logit form using the formula:

$$L_d = \log_e \left(\frac{p_d}{1 - p_d} \right) \quad (4)$$

and combine (4) with equation (2) to express L_d in terms of p_c:

$$L_d = \log_e\left(\frac{p_c - \frac{1}{3}}{1 - p_c}\right) \tag{5}$$

Substituting (3) into (5) gives the expression:

$$L_d = b(t - x) \tag{6}$$

Observe that L_d increases linearly with stimulus intensity if p_c conforms well to a logistic ogive. At this point, decide whether to complete the calculations using transformed or untransformed data. The untransformed graph of p_c versus log x is almost linear in its middle range, and for the purpose of locating the threshold, which is the point where $p_c = 2/3$, the transformation provides little advantage over an ogive fitted by eye or a straight line fitted numerically to the middle range of p_c data.

On the transformed scale, the threshold is the log concentration at which $L_d = 0$. Estimate its value from a straight line fitted visually or by fitting a linear regression line numerically. Note that transformed values near the asymptotes can be erratic because small changes in the proportions in these regions have large effects. Hence, ignore values of p_c below 0,43 or above 0,9 (transformed values of L_d below – 1,75 or above 1,75) when fitting the line to the plot, unless the interest of the experiment is in this region, see 6.5. Note also that the transformed graph permits a direct interpretation of the parameters of equation (3) as this is the stimulus intensity (in log concentration) at which $L_d = 0$ while b is the slope of the straight line.

6.3 Maximum likelihood procedure for fitting the data to a logistic model and estimating error bounds

6.3.1 General

The principle of the ML procedure for fitting the logistic model is to find those values of the parameters t and b for which the data are more likely than for any other values of the parameters. The procedure may be carried out and the error bounds estimated with the aid of one of the proprietary computer programs written for the purpose[2)], or perhaps more conveniently, by making use of the computer spreadsheet procedure described in Examples B.2 to B.4. The ML procedure finds, e.g., the upper bound as a value of t such that there is a probability of 0,05 of the estimate being greater than this.

6.3.2 The parameter b

This determines the steepness of the fitted line. For an individual assessor, it indicates fineness of discrimination for changes in stimulus intensity and is related to indices like the Weber ratio or the exponent in Stevens' power law. Individuals differ in b as they do in threshold and its value or distribution in the population may be of equal interest. Someone with a high value of b (a steep slope) is sensitive to small changes in intensity and might be particularly effective in tasks involving quality control or monitoring. Knowledge of an individual's b could be as important as of threshold when selecting assessors.

6.3.3 Confidence intervals for estimated parameters

The confidence interval for an estimate can be thought of as a range of values within which the true value might plausibly lie. The narrower the interval, the more confident we are about the estimate. The accuracy of the estimates can be improved by increasing the total amount of data and by choosing concentrations evenly spaced over a range of 0,25 × up to 4 × the threshold concentration.

---

2) GLIM [1] and SAS are examples. This information is given for the convenience of users of this International Standard and does not constitute an endorsement of these products.

© ISO 2002 – All rights reserved

6.4 Interpretation of results

Thresholds may be determined for a variety of purposes, and this document does not provide guidance on experimental design for particular purposes. When interpreting results and comparing thresholds, bear in mind how the data have been collected and analysed, and the degree of confidence to be placed in the derived statistics.

The results that are simplest to interpret and compare are those obtained for a single assessor. The fitted logistic model is the psychophysical function for the assessor and the derived statistics can be compared between assessors or between substances within assessors.

Data from different assessors or from different substances can be compared by extensions of the model for a single logistic function. Other designs may involve replication of presentations over several panels, representing different population groups. Comparison between substances or between panels for a given substance can be accomplished by standard ANOVA techniques using as the input data the estimates of t and b for the individual assessors providing that all the estimates were obtained in the same way, using the same number of presentations. If the data have been pooled as in 5.6.3.2, the resulting pooled t and b estimates can be used to describe differences between substances or between panels.

Distributions of thresholds over assessors may deviate widely from normality. Begin by examining the results in a histogram (see Figure 2) or in a normal or logistic probability plot. If skewness or bimodality is evident, calculate the appropriate average threshold, e.g. medians for skewed data, multiple averages for bi- or multimodal data.

Individual thresholds can be estimated from experimental designs in which each assessor makes at least one trial at each intensity, see Example B.1. The experimenter can then calculate the mean thresholds as well as the group threshold from pooled data, and examine the distribution of thresholds. Note that the individual thresholds estimated from such a design will be very imprecise while a pooled group threshold can have better precision. For example, a design incorporating two trials from each of 50 assessors at five intensities will provide 500 data points for estimation of the pooled threshold, but only 10 data points for estimating each individual threshold.

The group threshold from pooled data, and its error bounds, can be estimated from data derived from a design in which each assessor evaluates just one presentation at one concentration, but more often the design will require the assessor to make at least one trial at each concentration. Populations usually exhibit wide ranges of individual thresholds for a substance. For an individual, a 100-fold range of concentrations will typically span a range of p_d from 0,05 to 0,95, but the individual thresholds for a number of assessors can often span a 10 000-fold range of concentrations. A 100-fold range of concentrations presented to a group of assessors will mean that for some assessors, perhaps a substantial number, the entire test range will be near one of the asymptotes of that individual's psychophysical function. Results near the asymptotes make little contribution to estimating t and b, hence in the case of pooled data, individuals with high or low sensitivities will have low weight in estimating the parameters. If this is undesirable, pooling should not be used, or individuals of interest should be deliberately emphasized by asymmetrical weighting.

The data-fitting process assumes that the distribution of thresholds conforms to the logistic model, and any deviation from this distribution will show as a lack of fit of the data to the computed line. Lack of fit can be tested for using statistical procedures for goodness of fit, but it is unlikely that deviations from a single logistic model will be detected other than in experimental designs incorporating more than about 10 concentrations over more than a 500-fold range, and a total of a few hundred presentations over the range. If a test for goodness of fit reveals a significant lack of fit, models other than a single logistic function can be considered. The simplest will be the addition of a second logistic function with a different value for the t parameter, and possibly for the b parameter as well, for a proportion of the assessors. This will adequately model skewed and bimodal distributions.

6.5 p_ds other than 0,5

A regulator may wish to set a limit for a malodorous substance in air that will be detected on 5 % of occasions, or a flavourist may wish to determine the concentration of flavour added to a food that will be detected on 95 % of occasions that the food is tasted. These effect levels can be calculated from the logistic curve by finding the stimulus intensities corresponding to L_d values of – 2,94 and 2,94 respectively, with the required value of L_d being found from equations (4) or (5). If high or low values are to be determined, the investigator should ensure that there is an adequate amount of data in the region of interest so that the relevant intensities lie within the range for which data are obtained. Extrapolation beyond the range studied cannot safely be relied on.

6.6 Estimation of the Best Estimate Threshold (BET)

This shortcut procedure, see [2] and Example B.1, can be described as a risky and imprecise method of obtaining a rough estimate of a panel threshold. It is based on the threshold model, see 3.7. In the 3-AFC test there is a probability of 1/3 of making a correct selection at concentrations below the threshold, and a probability of 1,0 at concentrations above it. The procedure is economical in assessing time as only one presentation per concentration is made to each assessor. As a consequence, a larger number of assessors can be included.

Tabulate the data in ascending order of concentration (or in descending order of dilution, as in the example). Inspect the data for a complete run of successes as the concentration increases. Calculate the BET as the geometric mean of the highest concentration missed, and the next higher concentration. For example, in the case of assessor 1 the BET is $\sqrt{135 \times 45} = 78$.

This algorithm cannot be used when there is a complete run of correct selections or when there is an incorrect selection at the highest concentration, assessors 6 and 4 in the example. The recommended procedure is to continue testing at appropriate extended concentrations, but otherwise the following conventions can be adopted. If the selection at the highest concentration is incorrect, assume it would be correct at the next higher concentration in the sequence and calculate the BET accordingly. If there is a complete run of correct selections, assume the next lower concentration would be incorrect.

Calculate the BET for the group as the geometric mean of the individual BETs. A convenient measure of the variation between the assessors is the standard deviation of the $\log_{10}$ values, as in the example. BET results may be biased because the probability of a lucky guess is 1/3 and that of two or three lucky guesses in succession are 1/9 and 1/27. The procedure is risky because with only one presentation per assessor an above-threshold sample may be missed through confusion or inexperience with the stimulus or the mechanics of the test. The standard deviation of the log values may be underestimated if the BET falls near the extremes of the range of concentrations presented and if too few extended concentrations are tested.

6.7 Presentation of results

In reports of threshold tests, the following information shall be included:

a) all test conditions, such as the nature and source of the samples, the method of sampling, choice of medium (diluent), equipment and physical test set-up under which samples were presented to assessors;

b) concentrations or flowrates used, temperature and other conditions of the samples;

c) instructions and scoresheets given to the assessors;

d) dilution factor per step;

e) number of replications of the presentations per assessor;

f) composition of the panel with regard to age, gender and experience; additional information may be useful, e.g. familiarity with the stimulus evaluated, health, smoking, use of dentures, time since last meal, etc. No assessor shall be identified by name nor shall the report allow a reader familiar with the panel to refer a particular judgement to a particular panel member;

g) if the size of the data set is not very large, report the results as tables of numbers of presentations and of correct selections at the concentrations used, as in Table B.1. Report the estimated values of thresholds, as individual or group thresholds, and their error bounds;

h) if the distribution of thresholds in a population has been sought, report the individual thresholds and, as appropriate, derived statistics such as mean and variance, and measures of departure from a normal distribution;

i) if pooling has been resorted to, report the group threshold from pooled data and its bounds and also the slope t and its bounds. Note in the report that pooling of the (smaller) variance within an assessor with the (larger) variance between assessors may affect the calculation of the bounds.

© ISO 2002 – All rights reserved

Annex A
(informative)

Estimated number of assessors required for a given degree of precision

The threshold values obtained with the procedures described in this International Standard are highly uncertain because of the variability of human response to stimuli. Table A.1 shows examples from practice of the precision (± standard deviation) that can be expected in typical situations.

Table A.1 — Precision typically obtainable, as a function of panel size and number (n) of presentations

No.	Group tested	Purpose	Examples of number of 3-AFC tests presented	Examples of precision observed
1	One person	A physician wishes to know if the person is anosmic to substance X	1 presentation at each of 8 concentrations $n = 8$	± 2- to 5-fold
2	One person	In a bottler of flavoured water one wishes to know his/her sensitivity to substance X	6 presentations at each of 6 concentrations $n = 36$	± 50 % to 100 %
			10 presentations at each of 8 concentrations $n = 80$	± 20 % to 50 %
3	A panel of 8	An experimenter wishes to compare the sensitivity of two panels	1 presentation at each of 6 concentrations $n = 48$	± 1- to 3-fold
			6 presentations at each of 8 concentrations $n = 384$	± 20 % to 50 %
4	4 panels of 8 selected to represent sections of a population	A city engineer wishes to know the threshold level of substance X as a contaminant	1 presentation at each of 6 concentrations $n = 192$	± 50 % to 200 %
			6 presentations at each of 6 concentrations $n = 1\,152$	± 20 % to 50 %
5		The engineer wishes to determine the level of X undetectable by 95 % of the population	As above plus repeat tests with the most sensitive 25 % of the 4 × 8 persons n = approx. 1 800	± 40 % to 100 %

The relation between precision and number of 3-AFC tests is independent of the type of threshold sought, e.g. a maker of orange soda wishing to make certain that a particular ingredient can be detected by 95 % of the population would also need approximately 1 800 test presentations. The number required is highly dependent on the variability within the population, however, a population composed of widely different groups, e.g. many young and many old persons, would require much larger numbers for a given precision.

Annex B
(informative)

Examples

B.1 Example B.1: Shortcut procedure using BET — Odour threshold of a sample of chimney gas

B.1.1 General

The example illustrates the BET shortcut procedure of obtaining an approximate value near the threshold. The simplest form of the procedure is shown: a small number of assessors (nine) tested the six chosen concentrations only once, and no tests at extended concentrations was made for assessors 4 and 6. The correct procedure would have been to test, e.g., assessor 6 at dilutions of 3-, 9- and 27-fold higher, etc. until failure occurred and was confirmed. Improvement of the very poor precision could have been obtained by replicating the complete procedure.

B.1.2 Experimental

Six different concentrations of the chimney gas in odour-free air were prepared. Each of these was presented in conjunction with two samples of odour-free air. Concentrations were increased by a factor of three per concentration step. Nine randomly selected assessors from the population participated. Each proceeded from the lower to the higher concentrations, indicating at each step which sample was different from the other two. The results are presented in Table B.1.

Table B.1 — Example B.1: Odour threshold of chimney gas

Assessors	Dilution ratio (concentrations increase →)						Best Estimate Threshold (BET)	
	3 645	1 215	405	135	45	15	Value	Log_{10} of value
1	0	+	+	0	+	+	78	1,89
2	+	0	+	+	+	+	701	2,85
3	0	+	0	0	+	+	78	1,89
4	0	0	0	0	+	0	9	0,94
5	+	0	0	+	+	+	234	2,37
6	+	+	+	+	+	+	6 313	3,80
7	0	+	+	0	+	+	78	1,89
8	+	0	0	+	+	+	234	2,37
9	+	0	+	+	+	+	701	2,85
Sum of log_{10}s Average BET threshold = geometric mean Standard deviation of the log_{10} values							Σlog_{10} 209	20,85 2,32 0,81
"0" indicates that the assessor selected the wrong sample of the set of three; "+" indicates that the assessor selected the correct sample.								

© ISO 2002 – All rights reserved

B.1.3 Calculations

See [2]. BETs for each assessor were found as the geometric mean of the highest concentration missed and the next higher concentration. Assessor 4 was assumed to have been correct at the next lower dilution ratio of 1 to 5, and assessor 6 was assumed to have failed at the next higher dilution ratio of 1 to 10 935. The sum of the logarithms of the resulting BET values was found at 20,85 yielding an average of 2,32, the antilogarithm of which is the average threshold, a dilution ratio of 1 to 209. The standard deviation of the log values, a measure of the dispersion of thresholds, was calculated from the right-hand column. The value 0,81, indicates a large difference between assessors, but it is noted that the real value may be even larger as the range was curtailed by the experimenter's failure to test assessors 4 and 6 at extended concentrations.

B.2 Example B.2: Threshold from pooled data — Fitting of the ogive by the maximum likelihood procedure versus a least-squares procedure — Threshold of diesel taint in trout

B.2.1 General

The example illustrates the principle of the maximum likelihood procedure and presents a method of finding threshold parameters and bounds using a spreadsheet, applied to a large dataset of 18 3-AFC assessments at 9 unequally spaced concentrations, see B.2.3. For comparison, in B.2.4 a least-squares procedure is applied to the same data.

B.2.2 Experimental

A sample of diesel oil was tested for its potential to taint trout exposed to water containing the oil [5]. Trout were exposed to the diesel oil at several concentrations and after 24 h exposure were harvested and assessed against unexposed controls by a panel of 18 experienced assessors. Table B.2 summarizes the formulae used in spreadsheet format, and Table B.3 shows the experimental data and the worksheet for fitting the logistic model.

Table B.2 — Formulae used in the spreadsheet for estimating the parameters of the logistic, and the error bounds of the estimate of the threshold

Column	Formula
B	@LN(A3)
E	+D3/C3
F	(2/3)/(@EXP(G$18*(F$18-$B3))+1)+1/3
G	+$D3*@LN(F3)+($C3-$D3)*@LN(1-F3)
H	(2/3)/(@EXP(I$18*(H$18-$B3))+1)+1/3
I	+$D3*@LN(H3)+($C3-$D3)*@LN(1-H3)
J	(2/3)/(@EXP(K$18*(J$18-$B3))+1)+1/3
K	+$D3*@LN(J3)+($C3-$D3)*@LN(1-J3)

B.2.3 Results and calculations by maximum likelihood

The principle of the maximum likelihood procedure for estimating the parameters of the model is to find the values of the parameters that produce the maximum value of the likelihood function (equation B.1). Spreadsheet programs often contain functions which will find maxima or minima of functions and can conveniently be used to fit the logistic model to 3-AFC data. Apart from the fact that spreadsheets are widely available on personal computers in laboratories, and more so than statistical packages, there are some advantages in using spreadsheets rather than the statistical packages. The mechanics of the process are easy to understand for anyone familiar with the basic operations of a spreadsheet, and the operator does not need to acquire a knowledge of the programming

languages or conventions of sophisticated statistical packages. Finding the parameters and the bounds for the threshold proceeds smoothly with data sets with reasonable numbers of observations – more than 50 or so – which are spread around the threshold, but an advantage of the spreadsheet over statistical packages or specific programs is that the workings of the fitting procedure are more transparent to the operator and this can be an advantage when fitting less than ideal sets of data.

The commands for finding maxima or minima of functions are named differently in different spreadsheets and the example here uses the Optimizer function of the QuattroPro [3)] package. The equivalent in both Excel[3)] and Lotus 1-2-3 [3)] are Solver.

In Table B.3, columns A to D contain the experimental data. A has the concentration of diesel oil in the water (C) in ml/m^3, and column B the logarithms, in this case natural logarithms (although logarithms to other bases can be used) of the concentrations [ln(C)]. The stimulus intensities do not need to be in an equally-spaced geometrical series, or any other particular spacing. It is convenient, though not necessary, to sort them by concentration.

Columns C and D contain respectively the numbers of presentations (n) at each concentration and the corresponding numbers of correct selections (r). It is not a requirement that the numbers of assessors be the same at each concentration as here. Column E records the proportion of correct selections (P_{obs}) though these values do not figure in the subsequent calculations.

Columns F and G contain the calculations for fitting the logistic. The logistic is defined by the parameters b and t and for given values of the parameters the expected proportion of correct assessments can be calculated. The likelihood is the probability of obtaining the observed proportion assuming a model with the given values of the parameters. The joint probability over all the data sets is the product of the probabilities at each concentration. It is more convenient to work with the logarithm of the likelihoods, when the joint likelihood is the sum of the log likelihoods, and log probability is given by the expression:

$$\sum_{i=1}^{i=N} r_i \ln\left(P_{\text{est}}\right)_i + \left(n_i - r_i\right)\ln\left(1 - P_{\text{est}}\right)_i \qquad \text{(B.1)}$$

where N is the number of data sets and n, r, and P_{est} are the number of presentations, number correct and estimated proportion respectively. The likelihood is calculated in column G. Row 12 contains sums of the likelihoods, i.e. the logarithmic likelihood of the joint probabilities. Cell G12 has the formula + SUM (G3:G11) and is copied into cells I12 and K12.

The data are entered, in this example, in rows 3-11. Enter preliminary estimates of the threshold and slope parameter into cells F18 and G18 referenced in the formula in column F. The procedure is very robust to preliminary estimates well away from the optimum for sets of data approximately distributed around the threshold, and the preliminary estimates do not need to be close to the final values. The preliminary estimate of the threshold can be set at approximately the mid-point of the range of concentrations, and the slope parameter at 1 when natural logarithms are used in column B or 0,5 when logarithms to the base 10 are used. The optimizer or equivalent function is called, usually from the tools or equivalent menu. The function will need to be given the addresses of the variable cells, that is the cells that contain the values to be varied for maximization of the likelihood function, in this case cells F18 and G18, and the address of the solution cell, that is the cell that contains the value to be maximized, in this case G12. The function allows for various options, some of which will be set at default values. Check that the function is set to maximize the value in the solution cell. It is not necessary to set any constraints in this case. The function will have criteria for stopping the iteration based on the maximum number of iterations and the minimum change in the variables, and the operator may wish to reset these from the default values.

3) The spreadsheets mentioned are examples of suitable products available commercially. This information is given for the convenience of users of this International Standard and does not constitute an endorsement of these products.

© ISO 2002 – All rights reserved

Table B.3 — Example of the worksheet for fitting data to the logistic, and for calculating the error bounds of the estimate of the threshold using a spreadsheet program

	A	B	C	D	E	F	G	H	I	J	K
1						Optimum model		Lower bound		Upper bound	
2	C ml/m^3	ln(C)	n	r	P_{obs}	P_{est}	Likelihood	P_{est}	Likeli-hood	P_{est}	Likeli-hood
3	0,010	– 4,61	18	5	0,278	0,366	– 10,951	0,414	– 11,361	0,349	– 10,844
4	0,028	– 3,56	18	9	0,500	0,446	– 12,583	0,538	– 12,528	0,394	– 12,892
5	0,032	– 3,44	18	4	0,222	0,462	– 11,762	0,558	– 13,759	0,404	– 10,865
6	0,060	– 2,81	18	11	0,611	0,568	– 12,097	0,671	– 12,170	0,478	– 12,672
7	0,095	– 2,35	18	14	0,778	0,667	– 10,072	0,755	– 9,561	0,561	– 11,382
8	0,285	– 1,26	18	16	0,889	0,872	– 6,304	0,903	– 6,298	0,798	– 6,812
9	0,324	– 1,13	18	17	0,944	0,888	– 4,206	0,914	– 3,982	0,821	– 5,067
10	0,673	– 0,396	18	16	0,889	0,952	– 6,865	0,959	– 7,060	0,920	– 6,389
11	0,992	– 0,008 03	18	17	0,944	0,970	– 4,031	0,973	– 4,075	0,951	– 3,869
12							– 78,872		– 80,792		– 80,793
13											
14								Deviance	3,841		3,841
15											
16											
17						ln(t)	b	x	b	x	b
18						– 2,348 9	1,311	– 2,835 4	1,117	– 1,867 3	1,357
19					P						
20			Concentration units			0,095 44		0,058 70		0,154 5	
21											
22					Revised estimate of bound			– 2,835 4		– 1,867 3	

Start the iterations. The solution is obtained in a few seconds on a personal computer, perhaps with a statement that the iteration was stopped because the solution was converging too slowly. Cells F18 and G18 will now contain the required estimates of the parameters of the logistic.

In the present case, ln(t) = – 2,35, t = 0,095 ml/m^3, and b = 1,31.

These parameters define the logistic model which has the property of being the model with the maximum likelihood; this is referred to as the optimum model. Any other values of either parameter will give a model with a smaller likelihood. The difference in likelihoods between models is expressed as the deviance defined as:

$$\text{Deviance} = -2(l_1 - l_2)$$

where l_1 and l_2 are the log likelihoods for models 1 and 2. In this case, l_1 and l_2 are the optimum and alternative models respectively. The significance of the difference between models is determined by testing the deviance as χ^2 with the appropriate degrees of freedom. This relationship is used for finding the error bounds of the estimate of the threshold. The objective is to find a model with a value of t which gives a deviance equal to χ^2 with one degree of freedom at the selected confidence level. This is a deviance value of 3,841 for one degree of freedom and a confidence of p = 0,05.

The formulae in F3 and G3 can be copied into the corresponding columns in H to K if the appropriate column designators have been made absolute. Cell I14 has the formula + 2*($G12 – I12) for the deviance of the fitted model for the upper bound and the optimum model. Copy the formula into cell K14.

The bounds are obtained by an iterative process. It is necessary to start with a preliminary estimate of the bound. The size of the bounds is determined by several characteristics of the set of data, but by far the most important is the total number of observations in the data set. Table B.4 lists the size of step to be added, algebraically, to the estimated threshold to give a first estimate of a bound for various values of the total number of observations in the data set. In the example here, with a total of 162 observations, a first estimate of the lower and upper bounds would be – 2,84 and – 1,79. These are entered into cells H18 and J18.

Table B.4 — Step to be added to the estimated threshold to give preliminary estimates of the error bounds of the threshold

Total no. of observations	40	60	80	100	120	160	200	> 200
Step for upper bound	1,02	0,81	0,76	0,61	0,58	0,49	0,45	0,40
Step for lower bound	– 1,39	– 1,05	– 0,84	– 0,78	– 0,69	– 0,56	– 0,49	– 0,45

The optimizing function is used again, but with only the cell containing the b value, cell I18 or K18, as the variable. Cells I12 or K12 containing the sums, are the solution cells. If the value of the deviance is not adequately close to the target value, 3,84, a revised estimate of the bound is calculated. With data sets of more than 50 or so observations distributed around the threshold, the deviance is approximately linearly related to the square root of the step away from the threshold, and a revised estimate can be obtained by simple proportionality. Cell H22 contains the equations @SQRT(3,841/I14)*(H18-$F18)+$F18 to give a revised estimate of the bound, and is copied into cell J22. The revised estimate is entered, as a value, into H18 or J18 (or just edit the value there), and the optimizer is run again. The iteration is repeated until the deviance obtained is sufficiently close to the desired value. In the example of Table B.3, three iterations were enough.

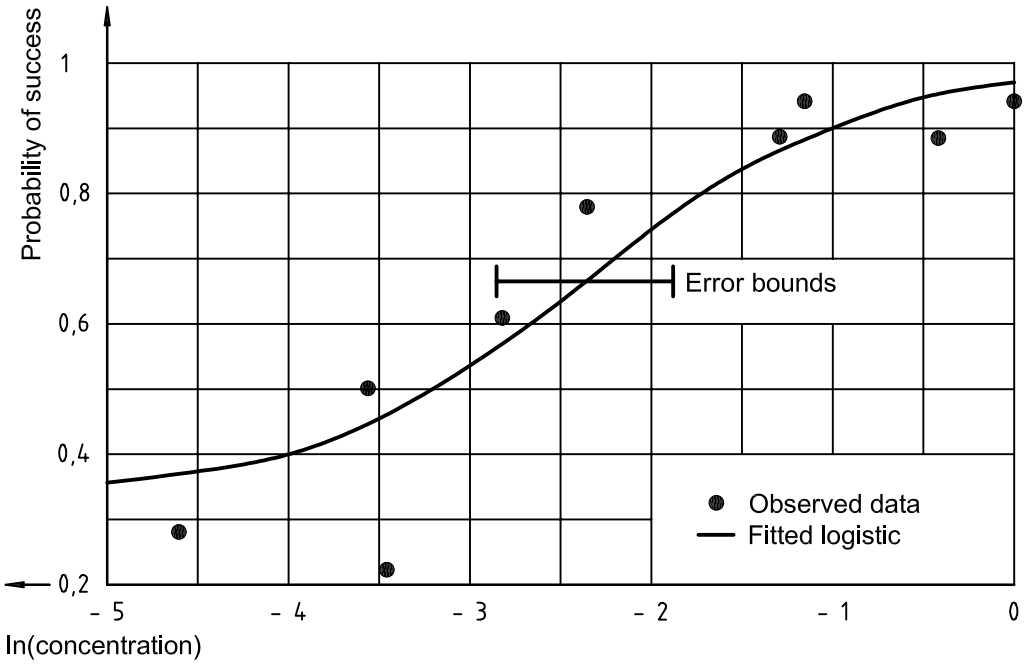


Figure B.1 — Example B.2: Group threshold from pooled data — Trout tainted by diesel oil

© ISO 2002 – All rights reserved

The operator can use the graphing facilities of the spreadsheet to plot the observed data and the fitted logistic, Figure B.1. The logistic can be plotted using the values in column H of Table B.3, but a smoother curve is obtained by constructing a vector of more closely, and evenly, spaced values of the logarithm of the concentration with the aid of the "fill" spreadsheet function and calculating the probabilities from the formula of column G of Table B.3 and the estimated parameters.

Difficulties in fitting the parameters, and particularly in finding the error bounds, will be experienced for ill-conditioned data sets. These will be small data sets or data sets which do not span the threshold, or in which values deviate markedly from a smooth monotonic increase in proportion of detections with increase in intensity of the stimulus. The optimizing function might fail to converge and might exit with a warning of some sort. It will usually be found that the program has exited with a bizarre value of b, very high, very low or even negative. (Where the numeric value of the experimental variable increases with increase in intensity, e.g. concentration of the stimulus, and the model requires that the probability of detection increases with increase in intensity, then b must be positive. When examining water or gases for off flavours or off odours it is common practice to dilute the sample and relate the proportion of detections to degree of dilution. In this case the numeric value of the test variable increases with decreasing intensity of stimulus and b will negative). The best option for the experimenter in cases of difficulty in fitting the logistic or finding bounds is to obtain more data, but otherwise the operator can repeat the fitting and include constraints on the value of b in the options of the optimizer function. The constraints could be a value of b greater than 0 or a low value, and less than 2 in the case of column B of Table B.3 expressed as natural logarithms, or 1 in the case of logarithms to the base 10. These upper constraints correspond to slopes such that the probabilities of detection in the range 0,05 to 0,95 are obtained over a 20-fold range of intensity of stimulus. If this fails to give a solution the operator can fix a value of b and fit only the threshold. This restricted fitting can be repeated over a range of values of b and the maximum likelihood function noted. An examination of the variation of the maximum likelihood with b will give an indication of whether or not an optimum solution exists.

The estimates of one or both bounds are likely to converge very slowly, or not at all, with ill-conditioned data using the algorithm of cell H22. An alternative approach is to record the values of t and the associated deviance in columns in the spreadsheet as the iteration proceeds. After three or four iterations, plot the deviance against the distance of the estimate from the threshold and extrapolate or interpolate visually for the value of the step at a deviance of 3,84. The visualization is made easier by subtracting 3,84 from the deviances and looking for the step where the corrected deviance is 0. The regression function of the spreadsheet can also be used to regress the reduced value of the deviance against the step as a second order polynomial. The resulting equation is solved to obtain t at a value of the reduced deviance of 0. This revised estimate of the bound is used in the next iteration. A convenient strategy is to replace the existing values of the reduced deviance and threshold in the vectors used in the regression furthest from the estimated bound by the new revised estimate of x and the corresponding deviance, and recalculate the expression to provide a further revised estimate. Again, the value of b in the optimizer function might have to be constrained to prevent its shooting off to unreasonable values. The process is repeated and soon converges if a reasonable value of the bound exists. Examination of the change of deviance with distance from the threshold will soon reveal if it will be possible to find a value for a bound.

B.2.4 Alternative calculation using linear regression

The principle of this procedure is to apply the logit transformation described in 6.2.2 and then use a proprietary computer program based on least squares regression such as SAS[4)] ProcReg to find the straight line that best fits the data. The commands for the data in Table B.3 and the resulting output are listed below.

4) This information is given for the convenience of users of this International Standard and does not constitute an endorsement by ISO of this product. Examples of equivalent products which will lead to the same results are those supplied by SPSS, S-PLUS and SYSTAT.

--`,,`,,`,`,,`,,,`,,,,`,,`,`,,`,`,,`---

```
TITLE "Logistic Regression of Threshold Data";
DATA INPUT;
  INPUT CONC R N;                 /*  Input Data                    */
  P=MAX(R/N, 1/3);                /*  Compute P(C)                  */
  LOGIT=LOG((P-1/3)/(1-P));       /*  Perform Logit Transformation  */
  LOG_C=LOG(CONC);                /*  Convert to LOG Concentration  */
CARDS;
.010  5 18
.028  9 18
.032  4 18
.060 11 18
.095 14 18
.285 16 18
.324 17 18
.673 16 18
.992 17 18
;
RUN;
DATA TRIMMED;                       /*  Trim data of extreme values  */
  SET INPUT;                        /*  as per Section 6.2.1         */
  IF LOGIT GE -1,75 AND LOGIT LE 1,75;
RUN;
PROC REG DATA=TRIMMED OUTEST=EST;       /* Fit the regression model  */
  MODEL LOGIT=LOG_C;
RUN;
DATA EST;
  SET EST;
  LOG_T    = -INTERCEP/LOG_C;           /* Compute Threshold value   */
  THRESHLD = EXP(LOG_T);
RUN;
```

© ISO 2002 – All rights reserved

```
PROC PRINT DATA=EST;                    /*  Output results          */
  VAR INTERCEP LOG_C LOG_T THRESHLD;
RUN;
                                                                      1
                        Logistic Regression of Threshold Data
Model: MODEL1
Dependent Variable: LOGIT
                        Analysis of Variance
                        Sum of          Mean
Source           DF     Squares        Square      F Value       Prob>F
Model             1     5,16632       5,16632       26,367       0,0143
Error             3     0,58783       0,19594
C Total           4     5,75415
Root MSE       0,44265     R-square       0,8978
Dep Mean       0,49539     Adj R-sq       0,8638
C.V.          89,35510
                        Parameter Estimates
                   Parameter      Standard    T for H0:
Variable  DF       Estimate         Error   Parameter=0    Prob > |T|
INTERCEP   1       2,368180    0,41498373         5,707        0,0107
LOG_C      1       0,900890    0,17544692         5,135        0,0143
-----------------------------------------------------------------------
                                                                      2
                   Logistic Regression of Threshold Data
             OBS     INTERCEP     LOG_C        LOG_T     THRESHLD
              1       2,36818    0,90089     -2,62871    0,072172
```

The resulting regression model is: LOGIT = 2,368 2 + 0,900 9 log(Conc). The threshold is the concentration at which LOGIT = 0, so the threshold concentration (in log units) is log(Conc) = – 2,368 2/0,900 9 = – 2,628 7 or (in standard units) Threshold = EXP(– 2,628 7) = 0,072 ml/m^3.

B.3 Example B.3: Individual thresholds — Comparing two assessors — Diesel oil in water

B.3.1 Objective

The objective was to determine if the two assessors differed in their perceptions of the substance. The example illustrates the use of a hierarchy of statistical models in the maximum likelihood procedure.

B.3.2 Experimental

The experiment used a chemical substance dissolved in water and the objective was to measure, and compare, the odour detection thresholds of two assessors.

B.3.3 Results and calculations

The approach [1] is to set up a hierarchy of statistical models of increasing complexity and to examine the likelihood that the data fit the models. As more parameters are added to the model the likelihood increases and the change in likelihood as parameters are added, expressed as the deviance, is examined to determine if the new model significantly increases the likelihood. In the case of fitting a logistic to 3-AFC data the parameters to be added to the model are values of t and b.

Table B.5 shows the experimental data and the calculations and Table B.6 contains a summary of the results obtained with three statistical models. The pattern of concentrations tested and the numbers of presentations at each concentration reflect the strategy of starting with a few presentations over a wide range of concentrations then conducting more replicate presentations at concentrations near the individual's expected threshold.

© ISO 2002 – All rights reserved

Table B.5 — Calculation of statistical models for comparing responses from two assessors

C ml/m^3	ln(C)	n	r	P_{obs}	Model 1		Model 2		Model 3	
					P_{est}	Log likeli-hood	P_{est}	Log likeli-hood	P_{est}	Log likeli-hood
Assessor 1										
0,001 5	– 6,50	3	0	0,000	0,337	– 1,23	0,334	– 1,22	0,334	– 1,22
0,004 0	– 5,52	3	0	0,000	0,347	– 1,28	0,335	– 1,23	0,335	– 1,23
0,004 6	– 5,38	3	0	0,000	0,350	– 1,29	0,336	– 1,23	0,336	– 1,23
0,013 9	– 4,28	3	0	0,000	0,404	– 1,55	0,348	– 1,29	0,349	– 1,29
0,020 0	– 3,91	5	1	0,200	0,442	– 3,15	0,360	– 2,81	0,360	– 2,81
0,025 0	– 3,69	7	1	0,143	0,473	– 4,59	0,371	– 3,78	0,372	– 3,78
0,041 7	– 3,18	3	1	0,333	0,565	– 2,24	0,415	– 1,95	0,415	– 1,95
0,100 0	– 2,30	13	8	0,615	0,759	– 9,33	0,578	– 8,70	0,578	– 8,70
0,125 0	– 2,08	3	3	1,000	0,804	– 0,65	0,636	– 1,36	0,636	– 1,36
0,500 0	– 0,693	5	4	0,800	0,961	– 3,41	0,926	– 2,91	0,925	– 2,90
2,500 0	0,916	2	2	1,000	0,995	– 0,01	0,994	– 0,01	0,994	– 0,01
12,500 0	2,53	2	2	1,000	1,000	0,00	1,000	0,00	1,000	0,00
Assessor 2										
0,001 5	– 6,50	3	0	0,000	0,337	– 1,23	0,344	– 1,27	0,344	– 1,27
0,004 0	– 5,52	3	2	0,667	0,347	– 2,54	0,384	– 2,40	0,383	– 2,40
0,004 6	– 5,38	3	1	0,333	0,350	– 1,91	0,395	– 1,93	0,394	– 1,93
0,012 5	– 4,38	8	4	0,500	0,395	– 5,73	0,560	– 5,60	0,559	– 5,60
0,013 9	– 4,28	3	2	0,667	0,404	– 2,33	0,586	– 1,95	0,585	– 1,95
0,020 0	– 3,91	5	4	0,800	0,442	– 3,85	0,682	– 2,68	0,682	– 2,68
0,041 7	– 3,18	3	2	0,667	0,565	– 1,97	0,855	– 2,25	0,856	– 2,25
0,050 0	– 3,00	7	6	0,857	0,604	– 3,95	0,886	– 2,90	0,886	– 2,90
0,100 0	– 2,30	5	5	1,000	0,759	– 1,38	0,958	– 0,22	0,958	– 0,21
0,125 0	– 2,08	3	3	1,000	0,804	– 0,65	0,970	– 0,09	0,970	– 0,09
0,500 0	– 0,693	5	5	1,000	0,961	– 0,20	0,997	– 0,02	0,997	– 0,02
2,500 0	0,916	2	2	1,000	0,995	– 0,01	1,000	0,00	1,000	0,00
12,50 0	2,53	2	2	1,000	1,000	0,00	1,000	0,00	1,000	0,00
Sums						– 54,48		– 47,76		– 47,76

Model 1: In this model one value each of t and of b is used for estimating values in the P_{est} column of Table B.5. The equations in the cells of this column reference two cells containing the values of t and b, and these two cells are used as the variable cells of the maximizing facility of the spreadsheet. The logarithmic likelihood is summed over the values for both assessors and this is contained in the target cell of the maximizing function. The function is run to give the estimates of t and b that maximize the summed likelihood. The values of t and b for this model are shown in Table B.6.

Table B.6 — Summary of the analysis of the models for comparison of the responses of two assessors

Model	Parameter	Assessor 1	Assessor 2	Log likelihood	Comparison	Deviance	D.F.	P, χ^2 test
1	t b	– 2,72 1,37		– 54,48				
2	t b	– 1,97 1,62	– 3,97	– 47,77	1 vs. 2	13,43	1	0,001 2
3	t b	– 1,97 1,63	– 3,97 1,62	– 47,77	1 vs. 3	0,00	1	

Model 2: An inspection of the data suggests that the thresholds for the two assessors are different and this model utilizes separate values of t for the assessors, but one value for b. The values in the P_{est} column of model 2 in Table B.5 are obtained using the appropriate values of t and the single value of b. The three cells containing the two values of t and the one value of b form the variable cells in the maximizing function, and the cell holding the log likelihoods summed over both sets of data is the target cell. The function is run to give the best values of t and b for the model. This time the model is a better fit to the data and the log likelihood is smaller. The values of the parameters and the deviance between this model and model 1 are shown in Table B.6.

Model 3: Model 2 is extended to include separate values of b for the two assessors. There are four variable cells, two for the ts and two for the bs, and, as before, the target cell is that holding the log likelihoods summed over both sets of data.

B.3.4 Summary and inferences

Table B.6 shows a summary of the parameters and the derived statistics. For each additional parameter added to a model in the hierarchy, one degree of freedom is lost, and, for these models, the deviance between two of them is tested as χ^2 with one degree of freedom. The deviance between models 1 and 2 is very highly significant with a probability of 0,001 2. The addition of a third parameter, the second b, to form model 3 does not improve the fit, a deviance of essentially zero.

It can be concluded that the two assessors differ in their odour detection thresholds for the diesel oil sample, but their sensitivity to increments in intensity does not differ.

B.4 Example B.4 : Comparison of thresholds of two substances. α- and β-pinene in water

B.4.1 General

The models applied in Example B.3 can equally well be applied to the comparison of two substances. The data from several assessors were pooled to yield a group threshold for each substance.

B.4.2 Experimental

The objective was to measure and compare the sensory properties of α- and β-pinene. The panel of 24 experienced assessors received one 3-AFC presentation at each of 10 concentrations.

B.4.3 Results and calculations

The experimental data and the spreadsheet calculations are shown in Table B.7. Table B.8 presents the results for t and b and a summary of the derived statistics.

© ISO 2002 – All rights reserved

B.4.4 Summary and inferences

Model 2 is a significantly better fit to the data than model 1 supporting the hypothesis that the thresholds for the two isomers are not the same. Model 3 in which two values of both t and b are fitted is an even better fit and has a larger likelihood. However, the deviance between models 2 and 3 of 3,37 has a probability of 0,066, not significant at the commonly used criterion of $p = 0{,}05$. It strongly suggests, however, that the slopes are not the same and this would warrant further investigation. In the case of β-pinene, the four lower concentrations are near the lower asymptote of the logistic and do not help to fix the slope. The concentrations for α-pinene cover the range between the asymptotes quite evenly and give a good estimate of the slope.

Additionally, the value of b in model 2 is much closer to that of α-pinene in model 3 again suggesting that the estimate of b for β-pinene is not as precise as that for the alpha isomer. Two, or more, further presentations of β-pinene at concentrations above the maximum used here, and perhaps one between 5 and 25 ml/m^3, would be advisable.

Table B.7 — Comparison of statistical models for calculating the thresholds of alpha and beta pinene

C ml/m^3	ln(C)	n	r	P_{obs}	Model 1		Model 2		Model 3	
					P_{est}	Log likeli-hood	P_{est}	Log likeli-hood	P_{est}	Log likeli-hood
α-pinene										
0,004	– 5,52	24	11	0,458	0,371	–16,93	0,387	– 16,80	0,412	– 16,66
0,020	– 3,91	24	8	0,333	0,408	–15,56	0,444	– 15,89	0,473	– 16,24
0,040	– 3,22	24	13	0,542	0,432	–17,13	0,480	– 16,73	0,508	– 16,61
0,100	– 2,30	24	14	0,583	0,473	–16,89	0,540	– 16,39	0,563	– 16,32
0,200	– 1,61	24	16	0,667	0,511	–16,46	0,593	– 15,55	0,609	– 15,45
0,500	– 0,693	24	15	0,625	0,571	–16,02	0,669	–15,98	0,673	– 16,00
1,000	– 2,30	24	20	0,833	0,622	–13,40	0,727	– 11,57	0,722	– 11,64
2,500	0,916	24	17	0,708	0,692	–14,50	0,797	– 15,02	0,782	– 14,85
5,000	1,61	24	19	0,792	0,743	–12,44	0,843	– 12,50	0,823	– 12,36
25,000	3,22	24	22	0,917	0,847	– 7,41	0,920	– 6,89	0,897	– 6,94
β-pinene										
0,004	– 5,52	24	11	0,458	0,371	– 16,93	0,349	– 17,16	0,334	– 17,35
0,020	– 3,91	24	10	0,417	0,408	– 16,30	0,368	– 16,42	0,336	– 16,64
0,040	– 3,22	24	12	0,500	0,432	– 16,86	0,381	– 17,33	0,338	– 17,96
0,100	– 2,30	24	8	0,333	0,473	– 16,24	0,407	– 15,55	0,346	– 15,28
0,200	– 1,61	24	11	0,458	0,511	– 16,69	0,433	– 16,58	0,358	– 17,06
0,500	– 0,693	24	9	0,375	0,571	– 17,74	0,479	– 16,40	0,393	– 15,89
1,000	– 2,30	24	8	0,333	0,622	– 19,35	0,523	– 17,02	0,445	– 15,90
2,500	0,916	24	13	0,542	0,692	– 17,73	0,591	– 16,67	0,558	– 16,57
5,000	1,61	24	18	0,750	0,743	– 13,50	0,648	– 14,07	0,672	– 13,85
25,000	3,22	24	21	0,875	0,847	– 9,12	0,779	– 9,77	0,894	– 9,09
Sums						– 307,21		– 300,32		– 298,63

Table B.8 — Summary of the analysis of the models for comparison of the thresholds of α- and β-pinenes

Model	Parameter	α-pinene	β-pinene	Log likelihood	Comparison	Deviance	D,F	P, χ^2 test
1	t	– 0,591		– 307,2				
	b	0,459						
2	t	– 0,723	– 1,831	– 300,3	1 vs. 2	13,79	1	0,000 20
	b	0,506						
3	t	– 0,787	– 1,581	– 298,3	1 vs. 3	17,16	2	0,000 19
	b	0,426	1,016		2 vs. 3	3,37	1	0,066

© ISO 2002 – All rights reserved

Bibliography

[1] AITKIN, M., ANDERSON, D., FRANCIS, B. and HINDE, J. Statistical Modelling in *GLIM*, Clarendon, Oxford 1989

[2] ASTM E679-91(1997), *Standard Practice for Determination of Odor and Taste Thresholds by a Forced-Choice Ascending Series Method of Limits*

[3] AMOORE, J.E. Specific anosmia and the concept of primary odors. *Chemical Senses and Flavor* **2**, 1977, pp. 267-281

[4] BUTTERY, R.G. Qualitative and sensory aspects of flavor of tomato and other vegetables and fruits. In *Flavor Science: Sensible Principles and Techniques*, ACREE, T.E. and TERANISHI, R., Chapter 8, pp. 277-278, ACS Professional Reference Book, American Chemical Society, Washington, DC, 1993

[5] DAVIS, H.K., GEELHOED, E.N., MACRAE, A.W. and HOWGATE, P. Sensory analysis of trout tainted by diesel fuel in ambient water. *Water Science and Technology* **25**(2), 1992, pp. 11-18

[6] DEVOS, M., PATTE, F., ROUAULT, J., LAFFORT, P. and VAN GEMERT, L.J. *Standardized Human Olfactory Thresholds*. IRL Press, Oxford 1990

[7] DOTY, R.L., GREGOR, T. and SETTLE, R.G. Influences of intertrial interval and sniff bottle volume on the phenyl ethyl alcohol olfactory detection threshold. *Chemical Senses,* **11**, 1986, pp. 259-264

[8] DOTY, R.L., DEEMS, D.A., FRYE, R., PELBERG, R. and SHAPIRO, A. Olfactory sensitivity, nasal resistance, and automatic function in the multiple chemical sensitivities (MCS) syndrome. *Arch. Otolaryngol. Head Neck Surg.,* **114**, 1988, pp. 1422-1427. (Description of the University of Pennsylvania's Dynamic Air-Dilution Olfactometer)

[9] EN 13725, *Air Quality — Determination of odour concentration by dynamic olfactometry*

[10] FAZZALARI, F.A. *Compilation of Odor and Taste Threshold Values Data*. American Society for Testing and Materials, Philadelphia, 1978

[11] GUADAGNI, D.G. and BUTTERY, R.G. Odor threshold of 2,3,6-trichloroanisole in water. *J. Food Science*, **43**, 1978, pp. 1346-1347

[12] VAN HARREVELD, A.Ph., HEERES, P. and HARSSEMA, H. A review of 20 years of standardization of odor concentration measurement by dynamic olfactometry in Europe. *J. Amer. Waste Management Assoc.*, **49** (5)

[13] MACMILLAN, N.A. and CREELMAN, C.D. *Detection Theory, A User's Guide*, Cambridge University Press, 1991, 395 pp

[14] van GEMERT, L.J. and NETTENBREIJER, A.H. *Compilation of Odour Threshold Values in Air and Water*. Central Institute for Nutrition and Food Research TNO, Zeist, Netherlands, 1977. Supplement V, 1984

ICS 67.240

Price based on 27 pages

© ISO 2002 – All rights reserved

--`,,`,,`,`,,,,,`,`,,,,``,,`,,`-`-`,,`,,`,`,,`---