INTERNATIONAL STANDARD

ISO 6658

Second edition
2005-10-01

Sensory analysis — Methodology — General guidance

Analyse sensorielle — Méthodologie — Lignes directrices générales

Reference number
ISO 6658:2005(E)

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 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.org
Web www.iso.org

Published 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 2.

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 document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 6658 was prepared by Technical Committee ISO/TC 34, *Food products*, Subcommittee SC 12, *Sensory analysis*.

This second edition cancels and replaces the first edition (ISO 6658:1985), which has been technically revised.

Introduction

This International Standard constitutes a general introduction to the methodology of sensory analysis and should be read before undertaking the more detailed test procedures described in other International Standards. It covers the general area of methodology and is intended to fulfil the following functions:

a) to provide a brief background of the essential features of methods of sensory analysis for the user of specific tests;

b) to provide details of general requirements, procedures and interpretation of results common to all or most tests;

c) to provide sufficient guidance on requirements, procedures and interpretation of results for the different specific tests to allow choice of the most appropriate procedure(s) for solution of a particular problem.

It comprises three main aspects, covered in Clauses 4, 5 and 6.

It is essential that Clause 4 "General requirements" be read first. Clause 5 "Methods of test" describes, in a general manner, all the main tests, under five headings:

— Definition;

— Application;

— Assessors;

— Procedure;

— Analysis of results.

Clause 6 is concerned with some general principles of data collection and analysis of sensory data and also briefly covers general principles of statistical treatment of the results.

Sensory analysis — Methodology — General guidance

1 Scope

This International Standard gives general guidance on the use of sensory analysis. It describes tests for the examination of foods by sensory analysis, and includes some information on the techniques to be used if statistical analysis of the results is required.

Generally these tests are intended only for objective sensory analysis. However, if a test can be used for determining preference, this is indicated.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 5492, *Sensory analysis — Vocabulary*

3 Terms and definitions

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

3.1
sensory analysis
examination of organoleptic attributes of a product by the sense organs

4 General requirements

4.1 Basic information

This clause covers the general requirements common to all situations encountered in sensory analysis. The information basic to these requirements is as follows.

a) The human response to one stimulus cannot be isolated from previous experience or from other sensory stimuli received from the environment.

NOTE Nevertheless, influences arising from these two sources can be controlled and the effect standardized.

a) Variability in sensory response is inherent in any group of people used for testing and is unavoidable; this can arise from inconsistencies within an individual, and through physiological and psychological differences between individuals.

NOTE However, with training, such a group can show highly consistent individual responses. Recognition of these factors is important in the analysis of results.

b) Systematic biases in sensory experiments involving human response can result in misleading data and incorrect interpretation that can be difficult to identify. The factors that can result in bias should be identified and controlled as far as possible by appropriate experimental design and conduct of the tests.

c) The validity of the conclusions drawn from the results is dependent upon the test used and the way it is conducted, including the questions that have been asked.

4.2 Statement of objectives

There are three main types of objective, as follows:

a) those in which the primary aim of the test is to categorize, rank or describe the product(s);

b) those in which the aim is to distinguish between two or more products; here it is important to distinguish between the need to know

- if there is a difference at all,
- how great is the magnitude of the difference,
- the direction (or quality) of that difference,
- the influence of that difference, e.g. with regard to preference, or
- if all or only part of a population is detecting a difference;

c) those in which reassurance is sought that products do not differ.

In sensory analysis, a given problem frequently requires appreciable discussion or thought before an appropriate test is selected. This is because the initial concept of the problem may require clarification.

4.3 Choice of test

The choice of appropriate test depends largely on the nature of the test objective, but also needs to take account of factors associated with the product, the assessors, the test environment, and the desired level of analytical precision and statistical confidence in the conclusions. The action that would occur based on the outcome of the test should be determined in advance.

For each test, an attempt is made in Clause 5 to give guidance as to its relevance. Preliminary tests may be necessary to confirm the applicability of a given test.

Because of sensory fatigue and the effects of adaptation, only a limited number of samples can be assessed during a session, depending on the nature of the test and the type of product. Some of these effects can be moderated by appropriate rinse procedures and recovery between samples.

Whilst the use of control samples is essential in most cases, their use naturally limits the number of samples that can be assessed during any given session.

The statistical plan should always be determined before commencing the tests. This is especially recommended if the number of samples to be evaluated requires more than one session. Details of statistical plans should be selected from specialized texts. Whatever test method is used, the sequential testing approach described in ISO 16820 should be considered whenever it is desirable to keep the number of samples or the number of assessors to a minimum.

4.4 Choosing and training assessors

A sensory analysis panel constitutes a true “measuring instrument”, and consequently the results of the analyses conducted depend on its members. The recruitment of persons willing to participate in a panel, therefore, needs to be carried out with care and should be considered as a real investment, both in time and financially. Management support in the organization is necessary if it is to be effective.

Sensory assessment may be made by three types of assessor: “assessors”, “selected assessors” or “expert assessors”. Assessors can be “naive assessors” who do not have to meet a precise criterion of selection or training, or people who have already taken part in some sensory tests (initiated assessor). Selected assessors are assessors who have been selected and trained for the particular sensory test. Expert assessors are assessors who have been selected and trained for a variety of sensory analysis methods and who demonstrate particular acuity in panel work.

NOTE Assessors employed by companies to undertake sensory analysis as their primary job function are examples of expert assessors.

The selection and training methods to be employed depend on the tasks and methods that it is intended to give to the selected assessors. Procedures for training assessors for descriptive tests are different from those for training assessors in discrimination tests.

Detailed procedures and methods for selection and training of assessors are given in ISO 8586-1. It should be noted that these methods sometimes only constitute a way of choosing the better candidates amongst those who are available, rather than to satisfy predetermined criteria. Also, the selection of assessors for their ability to discriminate and describe foods is quite different from that used for preference tests. The former tasks require selection and training, whereas the latter require only that the panel be representative of a specified sector of the population, for example, a group of consumers.

If a selection procedure is to be carried out, some important criteria for choosing assessors are as follows:

a) general ability to perform the specific sensory task, which may include a particular sensitivity to the stimuli under investigation;

b) availability with respect to normal employment;

c) motivation (willingness and interest);

d) good health (including the absence of specific allergies or treatment with medications) and good dental and general hygienic condition.

The performance of selected and expert assessors should be monitored regularly to ensure that the criteria by which they were initially selected continue to be met.

4.5 Material to be tested

The nature of the product to be tested determines the experimental protocol of the test, and may also have an influence on the type of test that is required to satisfy the test objectives. For example, a protocol in which foods are to be consumed hot will need to take into account the cooling rate of the product and the likely effect on sensory attributes, and the changes in sensory attributes that may occur in keeping the product hot prior to testing.

Methods of preparation and presentation of samples should be appropriate for the product and to the problem concerned.

EXAMPLE 1 A product that is normally consumed hot should be prepared in the usual manner and tested hot; however, elevated temperatures may be used in some circumstances to increase the ease with which some flavours can be evaluated.

EXAMPLE 2 A product that is normally consumed in discrete pieces should not be homogenized in order to retain textural characteristics. Care is needed, however, to ensure maximum uniformity between sub-samples for each assessor; this includes similar portion size and uniformity of composition.

General principles for product sampling (in accordance with International Standards relating to the product under test) should be applied for test samples. In all cases, documentation of sample identification codes or lot numbers is necessary. Valid conclusions can be drawn for a product as a whole only if the samples tested are representative.

Carriers may sometimes be used for tests relating to the evaluation of products for which direct tasting is not feasible (see ISO 5497), for example food ingredients.

Lighting conditions should be specified when appearance is being assessed. When the test concerns only differences in flavour, the effect of colour differences may be partially masked by the use of lighting conditions that minimize the colour difference.

Containers should be chosen so as not to affect the test or the product. These may include washable ceramic or glass containers, or disposable plastic or paper containers, but must not transfer chemical materials that could result in taint. In particular, washable containers should be washed only in odour- and taint-free detergents and rinsed in water, and polymeric and paper containers, including insulated containers used for hot or cold samples, should be odour- and taint-free.

Palate cleansers may be used by the assessors between samples and between sessions, but care should be taken to ensure that they do not influence the flavour of products to be assessed. Still and carbonated water and bland foods (for example, unsalted crackers) may be used between samples and between sessions. Checks on the water supply are desirable to ensure that it is bland. For particular purposes, deionized water, glass-distilled water, low mineral content spring water, carbon-filtered water or boiled tap water may be used, but it should be noted that they are likely to have different flavours.

4.6 Test room

Sensory analysis should be conducted in a dedicated test room (see ISO 8589 for details). The aim should be to create for each assessor a separate environment with minimum distraction, so that each assessor can quickly adjust to the nature of the new task(s). Extraneous activities, including preparation of the samples, should not be allowed during the tests, as these can lead to biased results. The room should be at a comfortable temperature and should be ventilated with odour-free air; limited airflow is desirable to avoid excessive temperature fluctuations. Persistent odours, such as tobacco or cosmetics, should not be allowed to contaminate the environment of the test room.

Sound should be restricted. A low background noise is usually more tolerable than a fluctuating level of noise. Conversation is more distracting than background noise. Interruptions cause the greatest distraction.

It is usually helpful to have control over both the colour and the intensity of the lighting, although coloured lights rarely succeed in completely masking differences in appearance.

Surfaces should be non-absorbent and designed to facilitate a high standard of hygiene. The dimensions of the tasting booths are important; very low ceilings and very narrow booths can be oppressive or can give rise to a feeling of claustrophobia. Comfortable seating is necessary.

If provision is made for computerized data acquisition, then this should be implemented safely, hygienically, and in a way that does not compromise sensory judgement.

4.7 Planning and conduct of the test

The planning and operation of the test are determined by the objectives of the programme, the test chosen, and practical constraints associated with the use of human subjects. In particular, it is important to recognize the biases that might be inherent in the chosen test, and to operate the test in such a way as to minimize the effects of any bias. The potential biases can originate from both psychological and physiological sources.

The most serious psychological bias results from assessors interacting to influence each other's judgements, and should be minimized by the use of individual booths or adequate separation of the assessors. Moreover, strict management of the activities of the assessors is necessary.

The manner and order of presentation of the samples are important aspects of the test, and can introduce psychological biases. For example, the samples should be coded by random three-digit numbers, and the codes should be varied for each test. The order of assessment can also be a source of bias and, in general, the order should be specified. With a small number of samples and assessors, the order can be balanced so that every possible order occurs an equal number of times. In larger experiments, the order can be balanced or randomized.

Physiological biases are frequently associated with the nature of the test samples. In particular, adaptation to a specific flavour stimulus can occur on repeated exposure to that stimulus, and fatigue can occur when chewing solid foods. Both factors can impose an upper limit on the number of samples to be assessed in a session. Expectoration of samples may be recommended with trained panels, but loss of information on specific sensory attributes may result.

Hunger and satiety can influence an assessor's performance, and, if panels are held too frequently, performance may deteriorate. If it is possible, assessors should be asked to refrain from smoking and from consuming snacks such as coffee for 1 h before a test. Assessors should not carry any foreign odours into the session, for example tobacco or cosmetic odours, as these could influence the responses of other assessors.

The time of day at which the test is conducted is important. The schedule should take into account local customary mealtimes since performance is generally considered optimum at mid-morning and mid-afternoon. Assessors suffering from emotional upsets, colds and other illnesses should be excluded from tests until they recover.

The collation of the results comprises three aspects:

— checking that all data have been recorded accurately, either on computer or manually;

— verification that any additional relevant information which may aid or cast doubt on the interpretation of the results has been noted;

— checking that the assessors are motivated to continue participating if further testing is planned.

5 Methods of test

5.1 General

The most commonly used tests are divided into three groups:

a) discrimination tests used to determine the probability of difference or similarity between products (see 5.2);

b) tests using scales and categories to estimate the order or size of differences or the categories or classes to which samples should be allocated (see 5.3);

c) descriptive tests used to identify the specific sensory attributes present in a sample (see 5.4).

For the number of assessors, refer to the corresponding standards, taking into consideration α or β risk depending on the purpose of the test. Alternatively, sequential analysis (see ISO 16820) may allow a decision to be made after fewer trials of the test than would be required by conventional approaches that use a predetermined number of assessments.

The tests may also be quantitative.

5.2 Discrimination tests

5.2.1 General

The following tests are commonly used to determine the probability of difference or similarity between samples:

a) paired comparison test (see 5.2.2);

b) triangle test (see 5.2.3);

c) duo-trio test (see 5.2.4);

d) two-out-of-five test (see 5.2.5);

e) “A — not A” test (see 6.2.6).

For all these tests, there are different manners of analysing the results.

5.2.2 Paired comparison test

5.2.2.1 Definition

This is a test in which samples are presented in pairs for comparison and detection of differences on the basis of some defined criteria.

See ISO 5495 for details.

5.2.2.2 Application

The paired comparison test is recommended

a) to determine if a perceptible difference exists in a particular attribute (e.g. sweetness), or to determine if no perceptible difference exists in that attribute,

b) to select, train and monitor the performance of the assessors,

c) to compare two products in terms of preference in the context of consumer tests.

Advantages of the test over other discrimination tests are simplicity and less sensory fatigue.

The disadvantage of the method of paired comparisons is that as the number of samples to be compared increases the number of inter-comparisons required rapidly becomes impracticable.

5.2.2.3 Procedure

The assessors receive a set of two samples (the pair). They designate the sample that they consider to be the most intense in the attribute under consideration, even if this choice is only a guess. One of the samples may be a control. The number of times each sample is selected is counted.

It is necessary to determine, prior to carrying out the test, whether the statistical test that follows is to be one-sided (i.e. the test supervisor expects a particular direction of difference and the alternative hypothesis corresponds to the existence of a difference in that direction) or two-sided (i.e. the test supervisor has no expected direction of difference and the alternative hypothesis corresponds to a difference in either direction).

Questions of difference and preference should not be combined: the recruitment criteria for panels are different for these questions.

5.2.2.4 Analysis of results

See 6.2.2.

5.2.3 Triangle test

5.2.3.1 Definition

This is a discrimination test involving three coded samples, two of which are identical, presented simultaneously. The assessors are asked to select the odd sample.

See ISO 4120 for details.

5.2.3.2 Application

The triangle test is recommended

a) when the nature of the difference is unknown, and

b) for the selection and training of assessors.

The test should not be used for the determination of preference. Some disadvantages of the test are that

— it is uneconomical for the assessment of a large number of samples,

— it may be more affected by sensory fatigue than the paired comparison test with intensely flavoured samples,

— if the nature of the difference is known, it is statistically less efficient than some other tests, and

— the method is applicable only if the products are as homogeneous as possible.

5.2.3.3 Procedure

The assessors are each presented with one set of three coded samples, two of which are identical, and are asked to select the odd sample.

Samples should be presented an equal number of times in each of the two sets of three distinct permutations of order, which are

BAA	ABA	AAB
ABB	BAB	BBA

5.2.3.4 Analysis of results

See 6.2.3.

5.2.4 Duo-trio test

5.2.4.1 Definition

This is a discrimination test in which the reference sample is presented first. It is followed by two samples, one of which is identical to the reference sample and which the assessors are asked to identify.

See ISO 10399 for details.

5.2.4.2 Application

This duo-trio test is used to determine if there is a sensory difference or similarity between a given sample and a reference. It is especially suitable when the reference sample is well known to the assessors, for example a sample of regular production.

If there are after-tastes, this test is less suitable than the paired comparison test (5.2.2) or the “A — not A” test (5.2.6).

5.2.4.3 Procedure

The assessors are first presented with the identified reference sample. This is followed by two coded samples, one of which is identical to the reference sample. The assessors are asked to identify this sample.

5.2.4.4 Analysis of results

See 6.2.4.

5.2.5 Two-out-of-five test

5.2.5.1 Definition

This is a discrimination test involving five coded samples, two of which are of one type and three of another. The assessors are asked to group the two sets of samples.

5.2.5.2 Application

The two-out-of-five test is recommended to establish a difference more economically than other tests (the method is statistically more efficient).

The disadvantages of this test are similar to those of the triangle test (5.2.3). It is more strongly affected by sensory fatigue and memory effects but has greater statistical power. Its principal use is in visual, auditory or tactile applications.

5.2.5.3 Procedure

The assessors are each presented with one set of five coded samples and are told that two are of one type and three of another. The assessors are asked to group the two sets of samples.

When the number of assessors is less than 20, the order of presentation should be selected at random from the following 20 distinct permutations:

> AAABB BBBAA AABAB BBABA ABAAB BABBA BAAAB ABBBA AABBA BBAAB ABABA BABAB BAABA ABBAB ABBAA BAABB BABAA ABABB BBAAA AABBB

5.2.5.4 Analysis of results

See 6.2.5.

5.2.6 “A — not A” test

5.2.6.1 Definition

This is a test in which a series of samples, which may be “A” or “not A”, is presented to the assessors after they have learnt to recognize sample “A”. The assessors are asked to indicate which sample is “A”.

See ISO 8588 for details.

5.2.6.2 Application

This test is a discrimination test, which may be used for the assessment of samples having variations of appearance or leaving a persistent after-taste.

It is especially useful when strictly similar repeated samples cannot be obtained.

5.2.6.3 Procedure

The assessors are presented with samples one at a time. The assessors are first presented with the reference sample “A” several times, until they can recognize it. They are then given several samples, each of which may be “A” or “not A”, at random, and they have to determine which they are. An appreciable time interval (for example 2 min to 5 min) should be allowed between receipt of samples, and only a few samples should be examined during one session.

5.2.6.4 Analysis of results

See 6.2.6.

5.3 Use of scales and categories

5.3.1 General considerations

See ISO 4121 for details.

Measurement methods in sensory analysis might seek to decide the categories, classes or grades to which samples should be allocated. They may also seek numerical estimates of the magnitude of attributes of samples or of differences between samples.

There is no direct relationship between the response scale used to elicit numbers and the measurement scale that corresponds to the values recorded. Thus the same method of obtaining numbers (response scale) can lead to values whose measurement scale is only ordinal (unequal intervals) or is on an interval scale (equal intervals). With an ordinal measurement scale, the size of the difference between two values cannot be assumed to reflect the difference between the perceived intensities. Nor can the ratio of two values be assumed to reflect the ratio of the perceived intensities. With an interval measurement scale, larger numerical values correspond to larger perceived intensities (or degrees of pleasure) and the size of the difference between two values reflects the size of the difference in perceived intensity of the property being measured. However, a numerical value of zero might not indicate a total absence of the property and the ratio of two values cannot be assumed to reflect the ratio of the perceived intensities.

The choice of response scale depends on the objectives of the study and the products being studied. In any specific case, there may be a choice among several equally good scales. Whatever response scale is adopted, it should be easy to use, discriminating, unbiased and easily understood by the assessors (see ISO 4121).

5.3.2 Interpretation of results

5.3.2.1 Quality of the measurements obtained

Irrespective of the response scale, the quality of the measurements depends on the manner in which they were obtained. Aspects to be considered are

— the training level of the assessors (see ISO 8586-1 and ISO 8586-2), and

— the method of presenting the samples (see 4.5 and 4.7).

5.3.2.2 Statistical power and interpretation

Statistical analysis is influenced by the nature of the measurement scale (ordinal, interval or ratio) rather than by the response scale used.

Results measured on an ordinal scale are best analysed using non-parametric methods, for instance, the Wilcoxon test in the case of two matched samples or the Friedman test with more than two samples. Measurements on an interval or ratio scale may be analysed by a parametric test, such as analysis of variance, if a normal distribution of residuals can be assumed.

Parametric tests are usually more powerful than non-parametric ones. That is, if a difference exists, the parametric test will be more likely to demonstrate it. On the other hand, non-parametric tests are more robust than parametric ones; that is, they are less affected by anomalies in the data.

In sensory analysis, the perception of a property is assessed, not the property itself, and it is impossible to be certain that equality of the intervals has been achieved. While it is not unusual to interpret the results as if they correspond to an interval or ratio measurement scale, this interpretation should be expressed in each specific case as a working hypothesis.

5.3.2.3 Types of measurement

It is useful to distinguish the following types of measurement:

a) classification (see 5.3.3);

b) grading (see 5.3.4);

c) ranking (see 5.3.5);

d) rating and scoring (see 5.3.6).

The term "measurement" is generic and refers to all of these.

5.3.3 Classification

5.3.3.1 General

Classification refers to a method of sorting samples (physically or by the labels that identify them) into predefined categories.

5.3.3.2 Application

Classification is applicable when it is wished to allocate samples to the most appropriate of several categories that are in no particular order. For instance, fish may be sorted according to their species, or samples may be allocated to categories according to the types of defect they possess. The principle is that each sample is allocated to the category of which it is most typical. If numbers are used to represent the categories, measurement is said to be on a nominal scale, with the numbers serving only as labels. For such numbers, neither the order nor the magnitude is informative.

5.3.4 Grading

5.3.4.1 General

Grading refers to a method of sorting samples into groups that constitute an ordinal scale of quality.

5.3.4.2 Application

Grading is applicable when it is wished to allocate samples to the most appropriate of several categories that are assumed to reflect quality. For instance, fish may be sorted according to freshness or samples may be allocated to categories according to the severity of their defects. The principle is that each sample is allocated to the category of which it is most typical. If numbers are used to represent the categories, measurement is said to be on an ordinal scale. For such numbers, only the order is informative.

5.3.5 Ranking

5.3.5.1 General

See ISO 8587 for details.

NOTE Compare with 5.3.6.

Ranking is a method of classification in which a series of samples is placed in order of intensity or degree of some specified attribute. No attempt is made to assess the magnitude of the difference between samples. If the position in the order is recorded as a number, the resulting measurement scale is ordinal.

5.3.5.2 Application

Ranking may be used as a rapid way of characterizing a small number of samples (about six) having complex attributes (for example, quality and flavour) or if a large number of samples (about 20) is to be assessed by appearance alone.

5.3.5.3 Procedure

It is necessary to ensure that the assessors understand and agree about the attribute or criterion on which the samples are to be ranked. Each assessor independently examines the coded samples in a prescribed order and assigns a preliminary ranking. The assessors should then review this ranking by re-examination of the samples and adjust it if necessary by rearranging the order.

5.3.6 Rating and scoring

5.3.6.1 General

Rating is a method of classification in which each sample is allocated to some point on an ordinal scale. More than one sample may be allocated to the same scale point. The scale may be numerical, verbal, graphic or a combination of these. It may be continuous or discrete and unipolar or bipolar (see ISO 4121). If the scale is numerical, the procedure is often called “scoring”. It may be useful for the assessors to have some samples as references to identify particular points on the scale.

5.3.6.2 Application

Rating may be used to evaluate the intensity of one or more attributes or degrees of liking of samples.

Although both ranking and rating invoke only ordinal scales, they are not equivalent. Ranking places the samples in order and its results consequently refer only to the group of samples ranked. Rating gives an ordinal estimate of the magnitude of attributes or preferences because the same ordinal scale is used irrespective of the particular samples being assessed. Thus rating is preferable if the results from one set of samples are to be compared with others. But since ranking encourages assessors to use any perceived differences among samples, it may reveal small distinctions among samples that would all be given the same rating.

5.3.6.3 Procedure

The classification to be used should be clearly defined and understood by the assessors. Each assessor independently examines the samples one-by-one in a prescribed order and assigns each to a point on a scale.

5.4 Descriptive tests

5.4.1 Types of test

These tests can be applied to one or more samples in order to characterize, both qualitatively and quantitatively, one or more sensory attributes. They may be classified as

a) simple descriptive tests (see 5.4.2),

b) sensory profile and descriptive analysis methods (see 5.4.3), or

c) free-choice profiling (see 5.4.4).

5.4.2 Simple descriptive test

5.4.2.1 Definition

This is a test to obtain a qualitative description of individual attributes contributing to the overall character of a sample.

5.4.2.2 Application

The test may be used

— for the identification and description of the attributes of a particular sample or samples, and

— for establishing the sequence in which these attributes are perceived.

The test is recommended for use in the description of previously established differences. It is a useful test for training assessors, or for preliminary vocabulary development for the more in-depth descriptive techniques described below.

5.4.2.3 Procedure

The test may be applied to one or more samples. When more than one sample is presented during a session, the order in which the samples are presented has an effect. The importance of this can be assessed by repeating the test, using a different order of presentation.

The sample is assessed independently by each assessor and the findings recorded. An attribute checklist may be provided. The sensory assessment may be followed by a discussion controlled by the leader of the panel.

5.4.2.4 Interpretation of results

The results should be collated to produce a list of descriptive terms applicable to the sample, based on frequency of usage of each descriptive word. Open discussion at the conclusion of the assessment is frequently helpful.

5.4.3 Quantitative descriptive and sensory profile tests

5.4.3.1 Definition

These are tests or theoretical methods of assessing, in a reproducible manner, the organoleptic properties of a product using terms selected from a glossary previously established by simple descriptive tests.

The separate attributes contributing to the overall sensory impression of the sample are rated on an intensity scale and the results are used to determine a sensory profile for the product. The method may be used for evaluating all sensations separately or in combination.

See ISO 6564, ISO 8586-1 and ISO 13299 for details.

5.4.3.2 Application

Descriptive tests are recommended

- for use in the development of new products,
- for establishing the nature of the differences between products,
- for use in quality control,
- to provide sensory data for correlation with instrumental data, and
- to provide sensory data for correlation with consumer data.

5.4.3.3 Assessors

A panel of selected assessors or expert assessors, specially trained in the method, is required.

According to the case

- the leader of the panel may be used to guide training,
- the leader of the panel may be used to control a discussion and to establish a consensus, or
- there may be no leader of the panel, but the situation is controlled by the organizer of the test.

5.4.3.4 Procedure

A preliminary set of trials (or training) is carried out with the range of products to be tested, so as to establish the organoleptic properties important in characterizing and distinguishing them. The results of these trials are used to develop a glossary of descriptive terms to be used, and to establish the experimental procedure for presenting and examining the samples. A panel is then trained in the methodology and particularly in the use of the glossary. It is useful at this stage to have a set of reference materials, pure compounds or natural products that elicit particular odour or flavour scores or have particular textural or visual properties.

In the test sessions, the assessors check the samples against the glossary of terms, scoring each attribute present on an intensity scale.

It is usual to note the order in which the factors are perceived, including the presence of an after-taste, and to score for the overall impression of aroma and flavour.

5.4.3.5 Interpretation of results

There are two basic approaches to the handling of data.

In consensus profile methods, immediately after the assessors have completed their assessments the leader of the panel tabulates the results and initiates a discussion to resolve differences. In the light of the discussion and, if necessary after re-examination of the samples, the panel arrives at a group decision on the profile.

In the other descriptive analysis methods, there is no discussion and the profile obtained is a series of averages of the scores assigned to each descriptor by each assessor.

The averages can be compared statistically, for example by using analysis of variance. There are also, for all descriptive analysis methods, techniques of multivariate analysis.

5.4.4 Free-choice profiling

5.4.4.1 Definition

This is a descriptive method in which untrained or minimally trained assessors evaluate products using their own individual sets of descriptors.

5.4.4.2 Application

These tests are recommended for use in the development of new products (particularly the perceptual mapping of product spaces). Their most distinct advantage is the avoidance of panel training.

5.4.4.3 Assessors

No specific training is necessary.

5.4.4.4 Procedure

Each panellist creates his or her own idiosyncratic list of descriptive terms by evaluating a broad range of samples and attempting to characterize and distinguish between them.

Assessors then conduct their individual evaluations of the test products using a traditional descriptive rating sheet constructed with their own vocabulary.

5.4.4.5 Interpretation of results

Data are generally submitted to Generalized Procrustes Analysis or Statis to produce a consensus perceptual map of important dimensions and to assess the degree of agreement of each assessor’s data to the statistically derived consensus map.

6 Analysis of results

6.1 General

This clause gives general indications of the appropriate methods to be used for the statistical analysis of the results of sensory tests. Further details on specific tests can be found in the appropriate International Standards listed in the Bibliography. Statistical terms in boldface are explained in Annex A and are in conformity with ISO 3534-1, ISO 3534-2 and ISO 3534-3.

6.2 Discrimination tests

6.2.1 General

The aim of the discrimination tests described in 5.2 is to determine if there is a detectable difference between two products, A and B (or a preference for one of them). The analysis is based on the number of assessors in each particular category, for example those preferring A, those preferring B, or those correctly choosing the odd sample.

The International Standard that deals in detail with each method also describes how to use it to give reassurance about similarity when that is required.

6.2.2 Paired comparison test (see ISO 5495)

6.2.2.1 Statistical interpretation

There are two possible forms of this test. The first is concerned with the detection and the determination of the direction of a specified difference between two products; the second is concerned with a preference for one of them.

In both cases, the **null hypothesis** is that no distinction can be made between the two products (either by intensity or by order of preference). Quantitatively, the null hypothesis is that there is an equal probability (1/2) that an assessor selected at random from the pool will select sample A or sample B.

The interpretation of results based on the number of participants designating A (or B) as having the greater intensity or being preferred, depends on the **alternative hypothesis** opposed to the null hypothesis. Depending on the nature of the alternative hypothesis, which is to be specified before carrying out the test, the test will be two-sided or one-sided.

6.2.2.2 Two-sided test

A two-sided test is one in which it is simply desired to find out if there is a difference in intensity between the two products (intensity test), or if either of the products is preferred to the other (preference test). The alternative hypothesis is written $P_A \neq P_B$ (i.e. either $P_A > P_B$ or $P_A < P_B$).

At a 5 % significance level, the null hypothesis is rejected if the number of votes for one sample is at least equal to that in column 2 of Table A.1.

If this is the case, the conclusion will be that there is a significant difference between the two products and, if the majority of votes is in favour of product A, the conclusion will be that, for the characteristic in question, A has a significantly greater intensity than B (or is significantly preferred, if that was the basis of the assessors' votes).

6.2.2.3 One-sided (directional) test

A one-sided test is one in which it is desired to discover if one of the specially designated products (A, for example) has a greater intensity than the other; the alternative hypothesis is then $P_A > 1/2$. A directional test is appropriate only if any outcome in the opposite direction would not be interpreted as a real effect but merely as a chance outcome that does not cast doubt on the null hypothesis.

At a 5 % significance level, the null hypothesis is rejected if the number of votes in favour of A is at least equal to that in column 4 of Table A.1. If this is the case, the conclusion will be that A's superiority over B (in intensity) has been significantly recognized by the panel.

6.2.3 Triangle test (see ISO 4120)

The null hypothesis is that it is not possible to distinguish between the products. In this case, the probability P of identifying the sample that is different from the other two is equal to $P_0 = 1/3$. In statistical terms, the null hypothesis H_0 is expressed by $P_0 = 1/3$.

The test is one-sided. The test supervisor wants to know if it is possible to distinguish between the two products, so will reject the null hypothesis in favour of the alternative hypothesis $P > 1/3$.

If the number of correct answers is greater than or equal to the corresponding figure in column 3 of the Table A.1, this corresponds to a proportion of correct answers significantly higher than $P_0 = 1/3$ at the 5 % significance level.

6.2.4 Duo-trio test

The null hypothesis is that it is not possible to distinguish between the products. In this case, the probability of identifying the sample that is identical to the reference sample is equal to $P_0 = 1/2$. In statistical terms, the null hypothesis H_0 is expressed by $P_0 = 1/2$.

The test is one-sided. The test supervisor wants to know if it is possible to distinguish between the two products, so will reject the null hypothesis for the alternative hypothesis $P > 1/2$ if the number of correct answers is greater than or equal to the figure in column 4 of the Table A.1 corresponding to the 5 % significance level.

6.2.5 Two-out-of-five test

The null hypothesis is $P_0 = 1/10$. The test is one-sided and the alternative hypothesis is $P > 1/10$. The number of correct answers is compared with the corresponding number in column 5 of Table A.1.

6.2.6 **“A — not A” test** (see ISO 8588)

The numbers of “A” responses and “not A” responses are totalled separately for the samples known by the sensory analyst to be “A” and for those known to be “not A”, giving a 2 × 2 table. A chi-squared (χ^2) test of independence or a Fisher “exact” test may be used to determine if the proportions of “A” and “not A” responses are different for the two types of sample.

The test is one-sided, with the null hypothesis that the two proportions are equal and the alternative hypothesis that the proportion of “A” responses is greater for the samples known to be “A”.

6.2.7 Treatment of “no difference” responses in discrimination tests

In discrimination tests, “no difference” responses can occur. It may, however, be stipulated that these are not permitted (the forced-choice technique). Doing so makes use of the responses of assessors who otherwise would have replied “no difference”, but whose judgements may nevertheless be predominantly correct. Its disadvantage is that it may antagonize assessors honestly wishing to record “no difference”.

If “no difference” results are permitted, the number of assessors reporting “no difference” is reported and the statistical analysis uses only the results from those who reported a difference. Conclusions are expressed as relating to assessors who express a preference or report a difference.

6.2.8 Systematic effects

Checks should be made to see if there are any systematic effects in relation, for example,

— to the order of testing for an individual assessor, and

— to the sequence of presentation to several assessors.

6.2.9 **Sequential approach** (see ISO 16820)

In the sequential approach, cumulative results from a discrimination test are examined continuously during the test. The test is stopped as soon as a decision can be reached. This approach often allows a decision to be reached after fewer trials than would be required using predetermined sample sizes.

6.3 Tests using scales and categories

6.3.1 General

See ISO 4121 for details.

The choice of a statistical method for sensory analysis using any of the tests indicated in 5.3 depends on the purpose of the test and on the number of products tested. This subclause gives information about the statistical methods used. For further details in the particular context of each test, relevant statistical textbooks should be consulted or advice sought from a statistician.

6.3.2 Classification

The results obtained for one type of product may be summarized as frequencies for each category. The chi-squared (χ^2) test may then be used to compare the distributions of two or more types of a product into the different categories, i.e., to test the null hypothesis that the distributions are the same in comparison with the alternative hypothesis that they are different.

6.3.3 Grading (see ISO 4121)

The data can be summarized as for classification. Results can alternatively be summarized by medians and products can be compared statistically by rank sum tests, although adjustments may be needed to allow for large numbers of tied ranks. If the data from a sample are used as a basis for a decision with respect to a larger quantity (a "lot"), see ISO 2859-1, ISO 2859-2, ISO 2859-3, ISO 2859-4 and ISO 3951 to obtain the characteristics of an appropriate sampling plan.

6.3.4 Ranking (see ISO 8587)

When the samples have been ranked by several assessors as indicated in 5.3.5, statistical tests may be carried out to determine if the samples are significantly different (rank sum tests). Tests may also be carried out to determine if a particular sample has a significantly higher or lower rank than the other samples.

The rank scores can be grouped into homogeneous clusters.

6.3.5 Rating

For rating on a discrete scale with a small number of points, the results for one sample may be treated as for classification. Continuous data, or discrete data with a large number of points, may be grouped and summarized by frequencies in each interval.

When more than one sample is rated, a non-parametric method should be used to compare the distributions obtained.

If the data satisfy the conditions of scoring, either as they stand or after transformation, then the methods given in 5.3.6 may be used.

6.3.6 Scoring

The results obtained for one sample may be summarized as a median or an average (arithmetic mean) with some measure of scatter (for example, range or standard deviation).

If only two samples are involved and the hypothesis of normality of the distribution of the scores is reasonable, a t-test may be used (see ISO 2854). If scores are obtained from more than two samples, the normal procedure is analysis of variance.

If the distribution of the scores from each sample seems not to be normal, it may be useful to use distribution-free methods.

6.4 Analytical or descriptive tests

See ISO 6564 for details.

Indications of the treatment of results for simple descriptive tests and for profile methods are given in 5.4.2.4 and 5.4.3.5.

Annex A
(informative)

Statistical terms

A.1 Test of a hypothesis — Null hypothesis

In sensory analysis, a test is carried out in order to evaluate a hypothesis. This hypothesis should be stated clearly and without ambiguity before performing the test. It should be formulated so that it can be translated into statistical language as a null hypothesis H_0.

The null hypothesis is generally the hypothesis that there is no difference between products in the intensity of a characteristic (or that there is no preference for one of them).

A.2 Alternative hypothesis

The alternative hypothesis is the clearly stated hypothesis that will be accepted if the null hypothesis is rejected. If the null hypothesis H_0 is $P = P_0$, the alternative hypothesis H_1 may be two-sided ($P \neq P_0$) or one-sided (for example $P > P_0$). Examples of both kinds are given in 6.2.2.

A.3 Significance and significance level

When the results of a test are analysed, there are two possible conclusions:

— the null hypothesis is not rejected;

— the null hypothesis is rejected.

As any test is performed by a limited number of assessors, the conclusion rejecting the null hypothesis (in favour of the alternative hypothesis) involves a risk. The significance level is the probability (or the maximum value of the probability) of rejecting the null hypothesis when the null hypothesis is true. This is called the “alpha risk”.

The classical logic of significance testing requires a decision to be made in advance about the acceptable alpha risk. Usually, the pre-assigned value for the significance level is $\alpha = 0{,}05$ (5 %) or $\alpha = 0{,}01$ (1 %). Most of the statistical tables used for interpreting test results include these two significance levels. It is important to note that the null hypothesis can be rejected “at the 5 % level” and not be rejected “at the 1 % level”.

If the null hypothesis is rejected “at the 1 % level”, it is effectively rejected also “at the 5 % level”. This explains why the expressions “significant” for the 5 % level and “very significant” for the 1 % level are sometimes used.

A.4 Mistaken conclusions: Type II — Efficiency

If the test does not lead to the rejection of the null hypothesis, this does **not** prove in any way that the null hypothesis is true. This only means that, from the limited information available (test with n assessors), there is insufficient reason to reject the null hypothesis (at the chosen significance level). The greater the amount of information (the larger n is), the more justified it is to reject the null hypothesis when it is false; the efficiency of the test increases with the number of assessors participating in the test. For example, in the case of a preference test (6.2.2) carried out with 20 assessors, the null hypothesis $P_0 = 1/2$ may not be rejected (the conclusion being that there is no significant preference for either of the two products), whereas if the test had been carried out with 100 assessors a significant preference for one of the products might have been demonstrated from the same proportions of the two choices.

The Type II error (which depends on the significance level chosen) is the probability (denoted by β) of mistakenly not rejecting the null hypothesis when the specified alternative hypothesis is actually true.

If the null hypothesis and the alternative hypothesis can be defined with the values of a parameter as in discrimination tests (paired comparison test, triangle test, duo-trio test, etc.), the Type II error can be calculated as a function of this parameter. For tests where the null hypothesis and the alternative hypothesis cannot be defined using the values of a parameter (evaluation tests, classification), it is generally not possible to calculate the Type II error.

Table A.1 — Significance tables

Number of assessors	Paired comparison test (two-sided)	Triangle test	Duo-trio test and paired comparison test (one sided)	Two-out-of-five test
5	—	4	5	3
6	6	5	6	3
7	7	5	7	3
8	8	6	7	3
9	8	6	8	4
10	9	7	9	4
11	10	7	9	4
12	10	8	10	4
13	11	8	10	4
14	12	9	11	4
15	12	9	12	5
16	13	9	12	5
17	13	10	13	5
18	14	10	13	5
19	15	11	14	5
20	15	11	15	5
21	16	12	15	6
22	17	12	16	6
23	17	12	16	6
24	18	13	17	6
25	18	13	18	6
26	19	14	18	6
27	20	14	19	6
28	20	15	19	7
29	21	15	20	7
30	21	15	20	7

NOTE All results refer to a significance level of 5 %. For other significance levels, see Reference [23].

Bibliography

[1] ISO 2854, *Statistical interpretation of data — Techniques of estimation and tests relating to means and variances*

[2] ISO 2859 (all parts), *Sampling procedures for inspection by attributes*

[3] ISO 3534-1, *Statistics — Vocabulary and symbols — Part 1: Probability and general statistical terms*

[4] ISO 3534-2, *Statistics — Vocabulary and symbols — Part 2: Statistical quality control*

[5] ISO 3534-3, *Statistics — Vocabulary and symbols — Part 3: Design of experiments*

[6] ISO 3951 (all parts), *Sampling procedures for inspection by variables*

[7] ISO 3972, *Sensory analysis — Methodology — Method of investigating sensitivity of taste*

[8] ISO 4120, *Sensory analysis — Methodology — Triangle test*

[9] ISO 4121, *Sensory analysis — Guidelines for the use of quantitative response scales*

[10] ISO 5495, *Sensory analysis — Methodology — Paired comparison test*

[11] ISO 5497, *Sensory analysis — Methodology — Guidelines for the preparation of samples for which direct sensory analysis is not feasible*

[12] ISO 6564, *Sensory analysis — Methodology — Flavour profile methods*

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

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

[15] ISO 8587, *Sensory analysis — Methodology — Ranking*

[16] ISO 8588, *Sensory analysis — Methodology — "A" — "not A" test*

[17] ISO 8589, *Sensory analysis — General guidance for the design of test rooms*

[18] ISO 10399, *Sensory analysis — Methodology — Duo-trio test*

[19] ISO 13299, *Sensory analysis — Methodology — General guidance for establishing a sensory profile*

[20] ISO 13300-1, *Sensory analysis — General guidance for the staff of a sensory evaluation laboratory — Part 1: Staff responsibilities*

[21] ISO 13300-2, *Sensory analysis — General guidance for the staff of a sensory evaluation laboratory — Part 2: Recruitment and training of panel leaders*

[22] ISO 16820, *Sensory analysis — Methodology — Sequential analysis*

[23] Roessler, Pangborn, Sidel and Stone, *J. Food Science*, **43**, 1978, p. 940

ICS 67.240

Price based on 20 pages