INTERNATIONAL STANDARD

ISO 8587

Second edition
2006-11-01

Sensory analysis — Methodology — Ranking

Analyse sensorielle — Méthodologie — Classement par rangs

Reference number
ISO 8587:2006(E)

PDF disclaimer

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 8587 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 8587:1988), which has been technically revised.

Sensory analysis — Methodology — Ranking

1 Scope

This International Standard describes a method for sensory evaluation with the aim of placing a series of test samples in rank order.

This method allows for assessing differences among several samples based on the intensity of a single attribute, of several attributes[1)] or of an overall impression. It is used to find if differences exist, but cannot determine the degree of difference that exists between samples.

The method is suited for the following cases:

a) evaluation of assessors' performance

   1) training assessors,

   2) determining perception thresholds of individuals or groups;

b) product assessment

   1) pre-sorting of samples

      i) on a descriptive criterion,

      ii) on hedonic preference;

   2) determination of the influence on intensity levels of one or more parameters (e.g. order of dilution, influence of raw materials, of production, packaging or storage methods)

      i) on a descriptive criterion,

      ii) on hedonic preference;

   3) determination of the order of preference in a global hedonic test.

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*

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

1) In this case, each attribute is tested through a different test in which the same products have different codes and are served in different orders to the same assessor.

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

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

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

ISO 3534-1, *Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability*

ISO 11035, *Sensory analysis — Identification and selection of descriptors for establishing a sensory profile by a multidimensional approach*

ISO 11036, *Sensory analysis — Methodology — Texture profile*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 3534-1 and ISO 5492 apply.

4 Principle

The assessors receive simultaneously three or more samples in random order.

NOTE Although it is possible to rank two samples, the paired comparison method, as described in ISO 5495[1], usually is preferred.

The assessors are asked to rank the samples according to a specified criterion: either a unidimensional criterion (i.e. particular attribute or specific characteristic of an attribute) or a global intensity (e.g. overall impression).

The rank sums are determined and statistical comparisons are made.

5 General test conditions

Refer, where available, to the standards (see ISO 6658) describing the sampling methods, the room in which tests are conducted (see ISO 8589) and the apparatus.

When preparing the test samples, the important points to be taken into consideration are as follows:

a) preparation, coding and presentation of the test samples;

b) number of samples to be compared that can be compared reliably (to be determined based on the nature of the test product (saturation sensitivity effects) and the design chosen; the number of samples shall be adapted based on

   1) the type of product [e.g. up to 15 samples can be assessed by selected assessors (ISO 8586-1) or experts (ISO 8586-2) on mild samples, while three can be a real maximum for harsh, spicy or high fat products assessed by consumers], and

   2) the criterion to be assessed (e.g. sweet is less saturating than bitter);

c) possible illumination of the samples.

6 Assessors

6.1 Qualification

The qualification of assessors depends on the aim of the test (see Annex A).

All assessors should preferably have the same level of qualification, this level being chosen according to the purpose of the test:

a) selected assessors or experts, for

   1) training assessors,

   2) assessment on a descriptive criterion, for instance determining the influence of intensity levels of one or more parameters (e.g. order of dilution, influence of raw materials, of production, packaging or storage methods),

   3) determination of perception thresholds of individuals or groups;

b) untrained assessors or consumers, already trained on the method

   1) for hedonic preference,

   2) when pre-sorting samples (to select a few products from a large number, as a preliminary test).

For the conditions with which assessors shall comply, see ISO 6658, ISO 8586-1 and ISO 8586-2. They shall all be specially trained on the ranking procedure and on the selected descriptors being used.

6.2 Number of assessors

The number of assessors depends on the aim of the test (see Annex A).

When testing assessors' performance, training assessors or determining perception thresholds of individuals or groups, no minimum or maximum number is required.

For descriptive product assessment, the minimum number of assessors is determined by the levels of statistical risks accepted and shall comply with ISO 11035 or ISO 11036, i.e. preferably around 12 to 15 selected assessors.

For determining the order of preference in a hedonic test, the minimum number of assessors is determined by the levels of statistical risks accepted, e.g. a minimum of 60 assessors per group of consumer type.

For statistical analysis of the results, other things being equal (for example, test conditions, qualification of assessors), the larger the number of assessors, the greater the probability of revealing any systematic difference in rank among products.

6.3 Preliminary discussion

The assessors shall be informed of the purpose of the test, i.e. ranking of test samples.

If necessary, a demonstration of a ranking procedure can be given. It is essential in this test to ensure common understanding by all assessors of the criterion under test. The preliminary discussion shall not influence assessors’ expectations.

7 Procedure

7.1 Presentation of the samples

The assessors shall not be able to draw conclusions about the samples from the way in which they are presented.

Prepare samples out of sight of the assessors and in an identical manner: same apparatus, vessels, same quantity of products, same temperature, same presentation. All irrelevant differences in samples shall be masked to avoid influencing the ranking. It is preferable to present the samples at the temperature at which the product is generally consumed.

The vessels are identified by 3-digit numbers, chosen at *random*, and different from one sample to another within one session (and preferably from one assessor to another).

The presentation takes into account the design chosen. In a “complete block” design, each assessor ranks all the samples. It is the preferred procedure. But, if the number of samples or their nature makes it impracticable to rank all the samples, a “balanced incomplete block” design may be used. In either case, it is necessary to ensure that all assessors complete their part of the design and do not omit any assessment.

For balanced incomplete block designs, each assessor is presented a specific subset of the samples in a randomized order (see Annex C for an example).

NOTE The use of a certain Balanced Incomplete Block is only possible when such a block variation exists in reality. So it is necessary to look for a predefined block from literature, e.g. Reference [5] in the Bibliography.

Each assessor is presented with k of the p samples ($k < p$). The subset of k samples is determined such that, in a single pass through the balanced incomplete block design, each sample is evaluated by n of the j assessors ($n < j$) and each pair of samples is evaluated by g assessors. It may be necessary to repeat the entire balanced incomplete block design several times in order to achieve an adequate level of sensitivity in the study. The number of repeats is denoted by r. In total, every sample is evaluated by $r \times n$ assessors and every pair of samples is evaluated by $r \times g$ assessors.

7.2 Reference samples

Reference samples may be included. If so, they are introduced unidentified into the series of samples.

7.3 Test technique

All assessors shall work under the same test conditions.

The assessors evaluate the samples presented in random order and place them in rank order on the designated attribute.

Instruct the assessors to avoid tied ranks[2]. If an assessor cannot differentiate two or more of the samples, instruct the assessor to place the samples in a rank order and record which samples they were unable to differentiate in the comment section of the answer form.

Provided there is no danger of sensory adaptation, and the products are sufficiently stable, it may be helpful to instruct each assessor to conduct an initial provisional ranking and then verify it by re-evaluating the samples in the rank order.

A single attribute shall be evaluated per test. If information about the ranking of more than one attribute is desired, each shall be evaluated by a separate test.

2) Tied ranks (identical ranks) are to be avoided and only used when assessors are really unable to differentiate between samples.

7.4 Answer form

An example of an answer form is shown in Annex D.

Sample codes should not initially appear on the blank answer form in case their positions influence assessors' expectations about their rank order. The ranks assigned to the individual samples shall be recorded by the assessors on the answer form.

Depending on the purpose of the test and on the test samples, it may be helpful to record additional information through a specific answer form.

8 Expression and interpretation of results

8.1 Summary of the results and calculation of the rank sums

Table 1 illustrates how rankings of one attribute by seven assessors for four samples are tabulated. If ranking is performed with respect to more than one attribute, a separate table is required for each attribute.

If there are tied ranks, record the mean rank of the samples that are tied. In Table 1, assessor 2 has assigned the same rank to samples B and C. Assessor 3 has assigned the same rank to samples B, C and D.

If there is no missing data and if tied ranks are correctly calculated, all rows will have the same total. The rank sum for each sample is obtained by adding up the ranks in each column. The rank sums indicate the consistency of the ranks assigned by the whole group of assessors. If they are consistent, the rank sums will be very different, but if they are inconsistent the rank sums will be similar.

8.2 Statistical analysis and interpretation

The statistical test to be chosen depends on the purpose of the test (see Annex A).

8.2.1 Determination of individual performance: Spearman correlation coefficient

To study the agreement between two rank orders (for example, rankings by two assessors or an assessor's rank order and an order predicted by information about the samples), the Spearman correlation coefficient, r_{S}, can be calculated:

$$r_{\mathrm{S}} = 1 - \frac{6\sum_i d_i^2}{p\left(p^2 - 1\right)}$$

where

p is the number of products ranked;

d_i is the difference between the two rankings for sample i.

If the value of the Spearman correlation coefficient approaches +1, there is high agreement between the two rank orders. If it is close to 0, the rank orders are unrelated.

If it approaches −1, there is strong disagreement between the rankings. Consideration should be given to the possibility that an assessor has misinterpreted the instructions and has arranged the samples in the opposite order to that intended.

Critical values of r_{S} to determine if the observed correlation is significant are given in Table 2.

8.2.2 Determination of group performance in the case of a predetermined order of samples, or confirmation of a predetermined order of samples: Page test [3]

This analysis can be used to determine if a panel of assessors collectively agrees with, or can perceive, the rank order of some property that a set of samples is known or predicted to have.

If $\Gamma_1, \cdots, \Gamma_p$ are the theoretical rank sums of the p samples in their predetermined order, the null hypothesis of absence of differences between the samples can be written: H_0: $\Gamma_1 = \ldots = \Gamma_p$

The alternative hypothesis is then: H_1: $\Gamma_1 \leqslant \ldots \leqslant \Gamma_p$, where at least one of these inequalities is strict.

For all products, the rank sums $R_1, \ldots, R_p$ are calculated (where R_1 is the rank sum for the sample that is first in the known rank order, and so on to R_p for the sample that is last in the known order).

To test the null hypothesis, H_0, calculate the Page coefficient L:

$$L = R_1 + 2R_2 + 3R_3 + \ldots + p \cdot R_p.$$

This coefficient will be highest when the theoretical ranking of products is reproduced by the assessors.

In the case of complete block designs, compare L with the critical values in Table 3, corresponding to the number of assessors, the number of samples and the chosen risk, for $\alpha = 0{,}05$ or $\alpha = 0{,}01$.

— If L is less than the tabulated value, no significant differences between the products are found.

— If L is equal to or greater than the tabulated value, there are significant differences between the rank sums of the products. H_0 is rejected and H_1 is accepted. It is concluded that the assessors tend to rank the samples in the predetermined order.

If the number of assessors or the number of samples is not in Table 3, calculate:

$$L' = \frac{12L - 3j \cdot p(p+1)^2}{p(p+1)\sqrt{j(p-1)}}$$

where

p is the number of products ranked;

j is the number of assessors.

This quantity approximately follows a standard normal distribution.

H_0 is rejected if $L' \geqslant 1{,}64$ (at the 0,05 risk) or $L' \geqslant 2{,}33$ (at the 0,01 risk) (see Table 3).

In the case of balanced incomplete block designs, calculate:

$$L' = \frac{12L - 3j \cdot k(k+1)(p+1)}{\sqrt{j \cdot k(k-1)(k+1)p(p+1)}}$$

where

p is the total number of products ranked;

k is the number of products ranked by each assessor;

j is the number of assessors.

Again, this quantity approximately follows a standard normal distribution.

H_0 is rejected if $L' \geqslant 1{,}64$ (at the 0,05 risk) or $L' \geqslant 2{,}33$ (at the 0,01 risk) (see Table 3).

Since the hypothesis H_0 was that all theoretical rank sums are equal, a significant result does not tell us that all sample differences are perceived, only that a difference between at least one pair of samples was consistently perceived in the predicted order.

8.2.3 Comparison of products where there is no assumed order

The Friedman test (Analysis of variance by ranks) [2] gives the maximum opportunities for demonstrating recognition by the assessors of differences among the samples.

8.2.3.1 Test if there is a difference between at least two products

This test is applied where j assessors have ranked the same p products.

Calculate the rank sums $R_1, R_2, \ldots, R_p$ of the p samples over the j assessors.

If $\Gamma_1, \cdots, \Gamma_p$ are the theoretical rank sums of the p samples, the null hypothesis of absence of differences between the samples can be written

$$H_0: \Gamma_1 = \ldots = \Gamma_p$$

The alternative hypothesis is that the rank sums for the population are not all equal.

For complete block designs, the Friedman test value is

$$F_{\text{test}} = \frac{12}{j \cdot p(p+1)}\left(R_1^2 + \ldots + R_p^2\right) - 3j(p+1)$$

where R_i is the rank sum of product i.

If $F_{\text{test}} > F$, from Table 4 considering the number of assessors, the number of products and the chosen risk, H_0 is rejected. It is concluded that there are consistent differences among the rank orders of the products.

For balanced incomplete block designs:

$$F_{\text{test}} = \frac{12}{r \cdot g \cdot p(k+1)}\left(R_1^2 + \ldots + R_p^2\right) - \frac{3r \cdot n^2 (k+1)}{g}$$

where

R_i is the rank sum of product i;

r is the number of repeats of the basic balanced incomplete block design;

k is the number of samples each assessor ranks;

n is the number of times each sample is evaluated in the basic balanced incomplete block design; and

g is the number of times each pair of samples is evaluated together in the basic balanced incomplete block design.

If $F_{\text{test}} > F$, from Table 4 considering the number of assessors, the number of products and the chosen risk, H_0 is rejected. It is concluded that there are consistent differences among the rank orders of the products.

If the number of samples or the number of assessors is not in Table 4, the critical values are found by an approximation that treats F_{test} as χ^2 with $p - 1$ degrees of freedom, where p is the number of products. Critical values of χ^2 are given in Table 5.

8.2.3.2 Test which products are significantly different from others

If it was concluded by the Friedman test that there are consistent differences among the rank orders of the products, then to determine which products are significantly different calculate a Least Significant Difference (LSD) at the chosen risk ($\alpha = 0{,}05$ or $\alpha = 0{,}01$).

In considering the level of α (the significance level, or risk of concluding that there is a difference when there is none), one of the two following approaches shall be chosen:

a) If the level of risk applies to each pair individually, then the risk to be associated is α. For instance, with a risk $\alpha = 0{,}05$ (i.e. 5 % risk), then, in the calculation of the LSD, the value of z (corresponding to a two-tailed normal probability of α) is 1,96. This is known as comparison-wise or individual risk. If the risk is α for each pair, there is a risk that is much greater than α of wrongly attributing a significant difference to one or more pairs in the whole experiment.

b) If the risk α applies to the whole experiment, then the risk to be associated with each pair of products is α', where $\alpha' = 2\alpha/p(p-1)$. For example, when $p = 8$, at risk $\alpha = 0{,}05$, then $\alpha' = 0{,}001\,8$ and then z (corresponding to a two-tailed normal probability of α') is 2,91. This is known as experiment-wise or global risk.

In most cases, it is the second of these, experiment-wise risk, that is most relevant to practical decisions about products.

For complete block designs:

$$\text{LSD} = z\sqrt{\frac{j \cdot p(p+1)}{6}}$$

For balanced incomplete block designs:

$$\text{LSD} = z\sqrt{\frac{r(k+1)(n \cdot k - n + g)}{6}}$$

If the observed difference between the rank sums of two products is equal to or greater than the LSD, then it is concluded that the two products have been given significantly different ranks.

If the observed difference is less than the LSD, then the two products have not been given significantly different ranks.

8.2.4 Tied ranks

If two or more ranks are tied, F for complete block designs is replaced by F':

$$F' = \frac{F}{1 - \left\{E / \left[j \cdot p\left(p^2 - 1\right)\right]\right\}}$$

where

E is obtained as follows:

Let $n_1, n_2, \ldots, n_k$ be the number of tied ranks in each group of tied ranks:

$$E = \left(n_1^3 - n_1\right) + \left(n_2^3 - n_2\right) + \ldots + \left(n_k^3 - n_k\right)$$

For example, in Table 1 there are two groups of tied ranks:

— the first group originates from assessor 2 (the two samples B and C are tied, thus $n_1 = 2$);

— the second group originates from assessor 3 (the three samples B, C and D are tied, thus $n_2 = 3$).

Hence:

$$E = \left(2^3 - 2\right) + \left(3^3 - 3\right) = 6 + 24 = 30$$

As $j = 7$ and $p = 4$, carry out the test, having calculated F, using the value:

$$F' = \frac{F}{1 - \left\{30 / \left[7 \times 4 \left(4^2 - 1\right)\right]\right\}} = 1{,}08\,F$$

Then compare F' with the critical values from Table 4 or 5.

8.2.5 Comparison of two products: Sign test

In the particular case of only two products being ranked, the sign test can be used.

NOTE In this case, the paired comparison test (ISO 5495) is a more appropriate test.

Where there are two products A and B, if k_A is the number of assessments in which product A was ranked first and k_B is the number of assessments in which B was ranked first, let k be the smaller number of k_A or k_B.

Any “no difference” responses should be disregarded.

The null hypothesis is

H_0: $k_A = k_B$ (A and B would be equally ranked in the whole population).

The alternative hypothesis is

H_1: $k_A \neq k_B$ (A and B would be differently ranked in the whole population).

If k is less than the critical value in Table 6 for the number of decisive assessments, H_0 is rejected and it is concluded that A and B have been given significantly different ranks.

9 Test report

The test report shall include the following information:

a) the aim of the test;

b) all information necessary for the complete identification of the sample (or samples)

   1) number of samples,

   2) whether reference samples have been used;

c) the test parameters adopted

   1) number of assessors and their level of qualification,

2) test environment,

3) material conditions;

d) the results obtained, together with their statistical interpretation;

e) reference to this International Standard (i.e. ISO 8587:2006);

f) deviations from this International Standard;

g) the name of the person supervising the test;

h) the date and time of the test.

Table 1 — Summary of the results and calculation of rank sums

Assessor	Samples				Rank sums
	A	B	C	D	
1	1	2	3	4	10
2	4	1,5	1,5	3	10
3	1	3	3	3	10
4	1	3	4	2	10
5	3	1	2	4	10
6	2	1	3	4	10
7	2	1	4	3	10
Rank sums for the samples	14	12,5	20,5	23	70

NOTE Since each assessor has assigned the same set of ranks, the row totals are identical and each is equal to $0,5 \times p(p + 1)$, where p is the number of samples.

Table 2 — Critical values for the Spearman correlation coefficient

Number of samples	Significance level (α)	
	$\alpha = 0,05$	$\alpha = 0,01$
6	0,886	—
7	0,786	0,929
8	0,738	0,881
9	0,700	0,833
10	0,648	0,794
11	0,618	0,755
12	0,587	0,727
13	0,560	0,703
14	0,538	0,675
15	0,521	0,654
16	0,503	0,635
17	0,485	0,615
18	0,472	0,600
19	0,460	0,584
20	0,447	0,570
21	0,435	0,556
22	0,425	0,544
23	0,415	0,532
24	0,406	0,521
25	0,398	0,511
26	0,390	0,501
27	0,382	0,491
28	0,375	0,483
29	0,368	0,475
30	0,362	0,467

Table 3 — Critical values for the Page test in the complete block design case

Number of assessors	Number of samples (or products) p											
	3	4	5	6	7	8	3	4	5	6	7	8
	Significance level						Significance level					
j	$\alpha = 0,05$						$\alpha = 0,01$					
7	91	189	338	550	835	1 204	93	193	346	563	855	1 232
8	104	214	384	625	950	1 371	106	220	393	640	972	1 401
9	116	240	431	701	1 065	1 537	119	246	441	717	1 088	1 569
10	128	266	477	777	1 180	1 703	131	272	487	793	1 205	1 736
11	141	292	523	852	1 295	1 868	144	298	534	869	1 321	1 905
12	153	317	570	928	1 410	2 035	156	324	584	946	1 437	2 072
13	165	343*	615*	1 003*	1 525*	2 201*	169	350*	628*	1 022*	1 553*	2 240*
14	178	368*	661*	1 078*	1 639*	2 367*	181	376*	674*	1 098*	1 668*	2 407*
15	190	394*	707*	1 153*	1 754*	2 532*	194	402*	721*	1 174*	1 784*	2 574*
16	202	420*	754*	1 228*	1 868*	2 697*	206	427*	767*	1 249*	1 899*	2 740*
17	215	445*	800*	1 303*	1 982*	2 862*	218	453*	814*	1 325*	2 014*	2 907*
18	227	471*	846*	1 378*	2 097*	3 028*	231	479*	860*	1 401*	2 130*	3 073*
19	239	496*	891*	1 453*	2 217*	3 193*	243	505*	906*	1 476*	2 245*	3 240*
20	251	522*	937*	1 528*	2 325*	3 358*	256	531*	953*	1 552*	2 360*	3 406*

NOTE Values marked with an asterisk (*) are critical values calculated by approximation using the normal distribution.

Table 4 — Critical values (F) for the Friedman test (risks of 0,05 and 0,01) [6]

Number of assessors	Number of samples (or products) p									
	3	4	5	6	7	3	4	5	6	7
j	Significance level $\alpha = 0{,}05$					Significance level $\alpha = 0{,}01$				
7	7,143	7,8	*9,11*	*10,62*	*12,07*	8,857	10,371	*11,97*	*13,69*	*15,35*
8	6,250	7,65	*9,19*	*10,68*	*12,14*	9,000	10,35	*12,14*	*13,87*	*15,53*
9	6,222	*7,66*	*9,22*	*10,73*	*12,19*	9,667	*10,44*	*12,27*	*14,01*	*15,68*
10	6,200	*7,67*	*9,25*	*10,76*	*12,23*	9,600	*10,53*	*12,38*	*14,12*	*15,79*
11	6,545	*7,68*	*9,27*	*10,79*	*12,27*	9,455	*10,60*	*12,46*	*14,21*	*15,89*
12	6,167	*7,70*	*9,29*	*10,81*	*12,29*	9,500	*10,68*	*12,53*	*14,28*	*15,96*
13	6,000	*7,70*	*9,30*	*10,83*	*12,37*	9,385	*10,72*	*12,58*	*14,34*	*16,03*
14	6,143	*7,71*	*9,32*	*10,85*	*12,34*	9,000	*10,76*	*12,64*	*14,40*	*16,09*
15	6,400	*7,72*	*9,33*	*10,87*	*12,35*	8,933	*10,80*	*12,68*	*14,44*	*16,14*
16	5,99	*7,73*	*9,34*	*10,88*	*12,37*	*8,79*	*10,84*	*12,72*	*14,48*	*16,18*
17	5,99	*7,73*	*9,34*	*10,89*	*12,38*	*8,81*	*10,87*	*12,74*	*14,52*	*16,22*
18	5,99	*7,73*	*9,36*	*10,90*	*12,39*	*8,84*	*10,90*	*12,78*	*14,56*	*16,25*
19	5,99	*7,74*	*9,36*	*10,91*	*12,40*	*8,86*	*10,92*	*12,81*	*14,58*	*16,27*
20	5,99	*7,74*	*9,37*	*10,92*	*12,41*	*8,87*	*10,94*	*12,83*	*14,60*	*16,30*
∝	5,99	*7,81*	*9,49*	*11,07*	*12,59*	*9,21*	*11,34*	*13,28*	*15,09*	*16,81*

NOTE 1 The quantity F may have only discontinuous values, this discontinuity being very pronounced for small values of j and p. Consequently, it is not possible to obtain critical values corresponding exactly to the risks 0,05 and 0,01.

NOTE 2 Values in italics were obtained using an approximation to the χ^2 distribution.

Table 5 — Critical values of χ^2 distribution (risks: 0,05 and 0,01)

Number of samples (or products)	Number of degrees of freedom of χ^2	Significance level α	
p	$(\nu = p - 1)$	$\alpha = 0{,}05$	$\alpha = 0{,}01$
3	2	5,99	9,21
4	3	7,81	11,34
5	4	9,49	13,28
6	5	11,07	15,09
7	6	12,59	16,81
8	7	14,07	18,47
9	8	15,51	20,09
10	9	16,92	21,67
11	10	18,31	23,21
12	11	19,67	24,72
13	12	21,03	26,22
14	13	22,36	27,69
15	14	23,68	29,14
16	15	25,00	30,58
17	16	26,30	32,00
18	17	27,59	33,41
19	18	28,87	34,80
20	19	30,14	36,19
21	20	31,4	37,6
22	21	32,7	38,9
23	22	33,9	40,3
24	23	35,2	41,6
25	24	36,4	43,0
26	25	37,7	44,3
27	26	38,9	45,6
28	27	40,1	47,0
29	28	41,3	48,3
30	29	42,6	49,6

Table 6 — Critical values for the sign test (two-tailed)

Number of assessments (j)	Significance level (α)		Number of assessments (j)	Significance level (α)	
	$\alpha = 0{,}05$	$\alpha = 0{,}01$		$\alpha = 0{,}05$	$\alpha = 0{,}01$
1			46	15	13
2			47	16	14
3			48	16	14
4			49	17	15
5			50	17	15
6	0		51	18	15
7	0		52	18	16
8	0	0	53	18	16
9	1	0	54	19	17
10	1	0	55	19	17
11	1	0	56	20	17
12	2	1	57	20	18
13	2	1	58	21	18
14	2	1	59	21	19
15	3	2	60	21	19
16	3	2	61	22	20
17	4	2	62	22	20
18	4	3	63	23	20
19	4	3	64	23	21
20	5	3	65	24	21
21	5	4	66	24	22
22	5	4	67	25	22
23	6	4	68	25	22
24	6	5	69	25	23
25	7	5	70	26	23
26	7	6	71	26	24
27	7	6	72	27	24
28	8	6	73	27	25
29	8	7	74	28	25
30	9	7	75	28	25
31	9	7	76	28	26
32	9	8	77	29	26
33	10	8	78	29	27
34	10	9	79	30	27
35	11	9	80	30	28
36	11	9	81	31	28
37	12	10	82	31	28
38	12	10	83	32	29
39	12	11	84	32	29
40	13	11	85	32	30
41	13	11	86	33	30
42	14	12	87	33	31
43	14	12	88	34	31
44	15	13	89	34	31
45	15	13	90	35	32

For values of j larger than 90, approximate critical values may be found by taking the nearest integer less than $(j-1)/2 - k\sqrt{j+1}$, where k is 0,980 0 for $\alpha = 0{,}05$ and 1,287 9 for $\alpha = 0{,}01$.

Annex A
(normative)

Determination of the test conditions

Table A.1 — Choice of the parameters of the test based on its aim

<table>
<tr><th rowspan="3">Test aim</th><th rowspan="3">Assessors qualification</th><th rowspan="3">Number of assessors</th><th colspan="3">Statistical method</th></tr>
<tr><th rowspan="2">Comparison to a known order
(assessors performance)</th><th colspan="2">Order of products unknown (products comparison)</th></tr>
<tr><th>2 products</th><th>> 2 products</th></tr>
<tr><td>Performance assessment of individuals</td><td>Selected assessors or expert sensory assessors</td><td>Unlimited</td><td>Spearman test</td><td rowspan="4">Sign test</td><td rowspan="4">Friedman test</td></tr>
<tr><td>Performance assessment of a group</td><td>Selected assessors or expert sensory assessors</td><td>Unlimited</td><td rowspan="2">Page test</td></tr>
<tr><td>Product assessment on a descriptive criterion</td><td>Selected assessors or expert sensory assessors</td><td>Preferably 12 to 15</td></tr>
<tr><td>Product assessment on hedonic preference</td><td>Consumers</td><td>Minimum 60 per group of consumer type (cell and segment)</td><td></td></tr>
</table>

Annex B
(informative)

Practical example of application — Complete block design

The results of fourteen assessors having tested one series of samples are compiled in Table B.1.

Table B.1 — Example of evaluation

Assessor	Samples				
	A	B	C	D	E
1	2	4	5	3	1
2	4	5	3	1	2
3	1	4	5	3	2
4	1	2	5	3	4
5	1	5	2	3	4
6	2	3	4	5	1
7	4	5	3	1	2
8	2	3	5	4	1
9	1	3	4	5	2
10	1	2	5	3	4
11	4	5	2	3	1
12	2	4	3	5	1
13	5	3	4	2	1
14	3	5	2	4	1
Rank sums	33	53	52	45	27

The value F_{test} from the Friedman test is calculated as follows.

Since $j = 14$, $p = 5$, $R_1 = 33$, $R_2 = 53$, $R_3 = 52$, $R_4 = 45$, $R_5 = 27$:

$$F_{\text{test}} = \frac{12}{14 \times 5 \times (5+1)}\left(33^2 + 53^2 + 52^2 + 45^2 + 27^2\right) - 3 \times 14 \times (5+1) = 15{,}31$$

The value of 15,31 is greater than that given in Table 4 for $j = 14$, $p = 5$ at the significance level of 0,05 (i.e. 9,32); it can therefore be concluded, with a risk of error less than or equal to 5 %, that the five samples have been perceived as being different.

Furthermore, it can also be decided that two individual samples are different if the absolute difference between their rank sums is greater than:

$$\text{LSD} = 1{,}96 \times \sqrt{\frac{14 \times 5 \times (5+1)}{6}} = 16{,}40 \text{ (at the 0,05 risk)}$$

At the 0,05 risk, the differences between A and B, A and C, E and B, E and C, E and D are significant, the differences between their rank sums being respectively:

A – B: $|33 - 53| = 20$ E – B: $|27 - 53| = 26$

A – C: $|33 - 52| = 19$ E – C: $|27 - 52| = 25$

E – D: $|27 - 45| = 18$

This last analysis could result in the following presentation:

E A D C B

The meaning of the underlining is as follows:

— two samples that are not connected by continuous underlining are perceived as significantly different (at the 0,05 risk);

— two samples that are connected by continuous underlining are perceived as significantly not different;

— A and E, which are not distinguished, are ranked significantly before D, C and B, which are themselves not distinguished. There are three groups present, one containing A and E, another containing A and D, the other B, D and C.

In the case where there are good reasons to assume *a priori* before testing that

rank (E) $\leqslant$ rank (A) $\leqslant$ rank (D) $\leqslant$ rank (C) $\leqslant$ rank (B),

then the Page test could be used to test this one-sided hypothesis.

Then, the L value from the Page test is calculated as follows:

$$L = (1 \times 27) + (2 \times 33) + (3 \times 45) + (4 \times 52) + (5 \times 53) = 701$$

The critical value from the Page test for $p = 5$, $j = 14$, $\alpha = 0{,}05$ is 661 (see Table 3).

As L is higher than 661, the null hypothesis of an overall absence of differences between the samples at the risk $\alpha = 0{,}05$ can therefore be rejected.

To conclude on this example:

a) based on the Friedman test

— at the 0,05 risk, E is not different from A; D is not different from C nor from B; A is not different from D, but A is significantly different from C and B, E is significantly different from D, from C and from B;

b) considering the Page test

— the assessors identify differences between samples at risk $\alpha = 0{,}05$; the assumed predetermined order is verified.

Annex C
(informative)

Practical example of application — Balanced incomplete block design

The results of ten assessors having tested three out of five samples from a balanced incomplete block design are compiled in Table C.1.

Table C.1 — Example of evaluation

Assessor	**Samples**				
	A	**B**	**C**	**D**	**E**
1	1	2	3		
2	1	2		3	
3	2	3			1
4	1		2	3	
5	2		3		1
6	1			3	2
7		1	3	2	
8		2	3		1
9		3		2	1
10			1	3	2
Rank sums	8	13	15	16	8

The value F_{test} for the Friedman test is calculated as follows.

Since $j = 10$, $p = 5$, $k = 3$, $n = 6$, $g = 3$, $r = 1$, $R_1 = 8$, $R_2 = 13$, $R_3 = 15$, $R_4 = 16$, $R_5 = 8$:

$$F_{\text{test}} = \frac{12}{1 \times 3 \times 5 \times (3+1)}\left(8^2 + 13^2 + 15^2 + 16^2 + 8^2\right) - \frac{3 \times 1 \times 6^2\,(3+1)}{3} = 11{,}6$$

The value of 11,6 is greater than that given in Table 4 for $p = 5$ at the significance level of 0,05 (i.e. 9,25); it can therefore be concluded, with a risk of error less than or equal to 5 %, that the five samples have been perceived as being different.

Furthermore, it can also be decided that two individual samples are different if the absolute difference between their rank sums is greater than:

$$\text{LSD} = 1{,}96 \times \sqrt{\frac{1 \times (3+1) \times (6 \times 3 - 6 + 3)}{6}} = 6{,}2 \text{ (at the 0,05 risk)}$$

At the 0,05 risk, the differences between A and C, A and D, C and E, D and E are significant, the differences between their rank sums being respectively:

A – C: $\lvert 8 - 15 \rvert = 7$	C – E: $\lvert 15 - 8 \rvert = 7$
A – D: $\lvert 8 - 16 \rvert = 8$	D – E: $\lvert 16 - 8 \rvert = 8$

This last analysis could result in the following presentation:

A E B C D

In the case where there were good reasons to assume *a priori* before testing that

rank(E) $\leqslant$ rank(A) $\leqslant$ rank(D) $\leqslant$ rank(C) $\leqslant$ rank(B),

then the Page test could be used to test this one-sided hypothesis.

The L value of the Page test is calculated as follows

$$L = (1 \times 8) + (2 \times 8) + (3 \times 16) + (4 \times 15) + (5 \times 13) = 197$$

Since $p = 5$, $k = 3$, $j = 10$, the L' value becomes

$$L' = \frac{12 \times 197 - 3 \times 10 \times 3 \times 4 \times 6}{\sqrt{10 \times 3 \times 4 \times 2 \times 5 \times 6}} = 2,4$$

As L' is higher than 2,33, the null-hypothesis of an overall absence of differences between the samples can therefore be rejected at the level $\alpha = 0,01$.

To conclude on this example:

a) based on Friedman test

at the 0,05 risk, A and E have significantly smaller rank sums than C and D. B is not significantly different from any of the other four samples;

b) based on the Page test

the assessors identify differences between the samples at risk 0,01. The assumed predetermined order is verified.

Annex D
(informative)

Example of an answer form

Name: ………………………… Date: ………………………..…. Test no: ……………………….

Please taste the samples from left to right.

Write the codes in increasing order of sweetness in the boxes below.

	Least			Most
Code				

Comments:

Bibliography

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

[2] FRIEDMAN, M. The use of ranks to avoid the assumptions of normality implicit in the analysis of variance, *Journal of the American Statistical Association*, **32**, 1937, pp. 675-701

[3] PAGE, E.B. Ordered hypotheses for multiple treatments: a significance test for linear ranks, *Journal of the American Statistical Association*, **58**, 1963, pp. 216-230

[4] XP V09-500, *Analyse sensorielle — Méthodologie — Directives générales pour la réalisation d'épreuves hédoniques en laboratoire d'évaluation sensorielle ou en salle en conditions contrôlées impliquant des consommateurs*

[5] COCHRAN, W.G. and COX, G.M. *Experimental Designs*. John Wiley & Sons, Inc., Chapter 11, Balanced Incomplete Blocks, 1950, pp. 315-346

[6] LOTHAR, S. *Applied Statistics, a Handbook of Techniques*, Springer Series in Statistics, Springer-Verlag, 1982

ICS 67.240

Price based on 21 pages