INTERNATIONAL STANDARD

ISO
8587
First edition
1988-12-01

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE DE NORMALISATION
МЕЖДУНАРОДНАЯ ОРГАНИЗАЦИЯ ПО СТАНДАРТИЗАЦИИ

Sensory analysis — Methodology — Ranking

Analyse sensorielle — Méthodologie — Essai de classement par rangs

Reference number
ISO 8587 : 1988 (E)

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.

Draft International Standards adopted by the technical committees are circulated to the member bodies for approval before their acceptance as International Standards by the ISO Council. They are approved in accordance with ISO procedures requiring at least 75 % approval by the member bodies voting.

International Standard ISO 8587 was prepared by Technical Committee ISO/TC 34, *Agricultural food products.*

Annex A forms an integral part of this International Standard. Annex B is for information only.

Sensory analysis — Methodology — Ranking

1 Scope

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

The method applies to performing multisample difference testing using the criterion of intensity of single attributes, components of attributes or total impression.

It is especially recommended for the pre-sorting of test samples (to be followed by the application of other test methods) or when other methods are beyond the capabilities of the assessors to use reliably.

Among other things, the method enables the influence of different raw materials, processing, treatment, packaging and storage to be determined.

It may also be suitable for use in the training of assessors.

2 Normative references

The following standards contain provisions which, through reference in this text, constitute provisions of this International Standard. At the time of publication, the editions indicated were valid. All standards are subject to revision, and parties to agreements based on this International Standard are encouraged to investigate the possibility of applying the most recent editions of the standards listed below. Members of IEC and ISO maintain registers of currently valid International Standards.

ISO 5492-1 : 1977, *Sensory analysis — Vocabulary — Part 1.*

ISO 5492-2 : 1978, *Sensory analysis — Vocabulary — Part 2.*

ISO 5492-3 : 1979, *Sensory analysis — Vocabulary — Part 3.*

ISO 5492-4 : 1981, *Sensory analysis — Vocabulary — Part 4.*

ISO 5492-5 : 1983, *Sensory analysis — Vocabulary — Part 5.*

ISO 5492-6 : 1985, *Sensory analysis — Vocabulary — Part 6.*

ISO 5495 : 1983, *Sensory analysis — Methodology — Paired comparison test.*

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

ISO 8589 : —[1], *Sensory analysis — General guidance for the design of test rooms.*

3 Definitions

See ISO 5492 for the definition of terms relating to sensory analysis. Statistical terms are used as defined in accordance with ISO 3534 : 1977, *Statistics — Vocabulary and symbols.*

4 Principle

Simultaneous presentation of several test samples in random order to assessors.

Ranking of the samples according to a specified criterion (for example, total impression, particular attribute or specific characteristic of an attribute). If a reference sample is used, it is placed unidentified among the other samples.

Statistical evaluation of the test results.

5 Apparatus

The apparatus shall be selected by the test supervisor, according to the nature of the product, the number of samples, etc., and shall in no way affect the test results.

If standardized apparatus corresponds to the needs of the test, it shall be used.

1) To be published.

6 Sampling

Refer to the standards relating to sampling, for sensory analysis, of the product or products to be examined.

The method of sampling shall take account of the test objective and, if there is no standard for the product concerned, shall be agreed between the interested parties.

7 General test requirements

7.1 Room

For the design of, and the conditions in, the room in which tests are to be conducted, see ISO 8589.

7.2 Assessors

7.2.1 Qualification

For the conditions which the assessors shall fulfil, see ISO 6658.

All the assessors shall preferably have the same level of qualification, this level being chosen according to the purpose of the test (for example, untrained assessors, selected assessors or experts).

7.2.2 Number of assessors

The number of assessors depends on the aim of the test. In general *at least* five selected assessors are necessary. For statistical analysis of the results, the larger the number of assessors, the greater the probability of revealing differences in rank between products.

All assessors shall work under the same test conditions.

7.3 Preliminary discussion

The assessors shall be informed of the purpose of the test. If necessary, a demonstration can be arranged on the basis of ranking. It is essential in this test to ensure common understanding of the criterion tested by all assessors. The preliminary discussion shall not influence the test results.

8 Preparation of test samples

Where applicable, specify in accordance with the purpose of the test

— the kind of preparation and the presentation of the sample;

— the number of samples;

— the temperature of samples (identical for all samples in the same test);

— any masking of single attributes (for example the use of coloured lamps to remove the effects of colour; homogenizing of samples to remove the effects of texture).

For each sample, the quantity of product presented to the assessor shall be the same and shall be sufficient to allow an evaluation of each of the samples as many times as required.

The number of samples to be ranked will be determined in accordance with the degree of difficulty of the test. (For example, for assessing samples of intense flavour, the number shall be kept small; on the other hand, for ranking according to the criterion "colour", the number of samples may be greater.)

The vessels containing the test samples shall be coded, preferably using three-digit random numbers. The coding shall be different for each test.

9 Procedure

9.1 Presentation of samples

The assessors shall not be able to draw conclusions as to the nature of the samples from the way in which they are presented. Therefore, the various samples of a series shall be presented in an identical fashion (same apparatus, same vessels and same quantities of products).

9.2 Reference samples

Reference samples may be introduced. In this case, these samples are introduced unidentified into the series of samples.

9.3 Test technique

The assessors evaluate P samples presented in random order and place them in a certain rank order, restricting themselves to the specified criterion.

The same series of samples may be presented to each assessor one or more times with different codes.

As instructed by the test supervisor, the assessors assign the rank 1 to the sample with the strongest or weakest intensity of the attribute to be tested (for example, hardest/softest, sweetest/least sweet). The ranks 2 to P are given to the other samples in sequence.

The assessors should be instructed to avoid tied rankings. They should be told that, even when they cannot find much difference between two samples, they should make a best guess.

However, if they are unable to differentiate between samples, they shall make a note of this on the answer form (see 9.4) in the section reserved for comments.

It is recommended that assessors set up a provisional rank order first and then verify it by further evaluation into order of increasing intensity. Only one attribute may be considered. If a rank order for several attributes is required, the test supervisor shall set up a separate experiment for each attribute.

Instructions specific to certain products shall be indicated (for example, "stir before assessing the odour"). In the case of gustatory stimuli, the assessors may be invited to use auxiliary products, such as water, weak tea or white bread, to neutralize sensations between assessments.

9.4 Answer form

The ranks of the individual samples shall be recorded on an answer form. A specimen answer form is given in annex B. Depending on the purpose of the test and the test samples, it may be necessary to record further information.

10 Expression and interpretation of results

10.1 Collation of results

If necessary, summarize in a table the evaluations recorded on the answer forms by each of the assessors, for each test and for each attribute, indicating tied ranking samples by an "equal to" sign (see table 1).

Table 1 — Summary of assessors' evaluations

Assessor	Rank order			
	1	2	3	4
1	A	B	C	D
2	B	= C	D	A
3	A	B	= C	= D
4	A	D	B	C
5	B	C	A	D

NOTE — For convenience, the letters A, B, C and D have been used in tables 1 and 2. In table 1, the letters represent the coded samples using random three-figure numbers as presented to assessors. In table 2, the letters represent the actual names of the samples.

10.2 Decoding of the samples and calculation of the rank sums

Decode the samples and tabulate the rank orders given to each sample by each assessor. Where there are tied rankings, note the mean rank (see table 2). Calculate the rank sum for each sample on the basis of the results by adding column sums in table 2.

By comparing the rank sums for the samples it is possible to obtain an evaluation of the differences between the samples.

Table 2 — Decoding and calculation of rank sums for the example given in table 1

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
Rank sums for the samples	10	10,5	13,5	16	50

NOTE — The row totals are identical and equal to $0,5\,P\,(P + 1)$.

10.3 Statistical interpretation

From the numerous tests featured in the literature, the following are recommended:

— the Friedman test[1], which, in a very general way, gives the maximum opportunities for demonstrating recognition by the assessors of differences between the samples[2]; and

— the Page test[3], where there is a pre-determined order of the samples.

---

1) 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.

2) In response to the question, "Are there differences in assessment between the samples?", other techniques are also available:

— the technique proposed by Fisher, which consists simply of carrying out an analysis of variance after first converting the ranks into scores;

— two other techniques which are based on the rank sum:

  — the Kramer test (not recommended): this test, in current use, has been shown to be better only in the case of a very small number of samples being different from the rest [moreover, it has been shown that the most widely used tables for this test are incorrect (see clause A.2)];

  — the test of the difference of extremes (also called "Wilcoxon multiple comparison method").

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.

10.3.1 Friedman test

10.3.1.1 Overall comparison of all samples

Calculate the Friedman value F, as follows:

$$F = \frac{12}{JP\,(P + 1)}\,(R_1^2 + R_2^2 + \ldots + R_P^2) - 3J(P + 1)$$

where

J is the number of assessors;

P is the number of samples (or products);

$R_1, R_2, \ldots, R_P$ are the rank sums attributed to the P samples for the J assessors.

Then compare F with the critical values from table 3.

If F is equal to or greater than the critical value corresponding to the number of assessors, the number of samples and the selected level of significance of $\alpha = 0{,}05$ (5 % level) or $\alpha = 0{,}01$ (1 % level), it can be concluded that there is a significant overall difference between the samples.

As the number J of assessors becomes greater, F follows approximately the χ^2 distribution with $(P - 1)$ degrees of freedom [values marked with a double asterisk (**) in table 3].

Table 3 — Approximate critical values from the Friedman test (levels of 0,05 and 0,01)

Number of assessors J	Number of samples (or products) P					
	3	4	5	3	4	5
	Level of significance $\alpha = 0{,}05$			Level of significance $\alpha = 0{,}01$		
2	—	6,00	7,60	—	—	8,00
3	6,00	7,00*	8,53	—	8,20*	10,13
4	6,50	7,50*	8,80	8,00	9,30*	11,00
5	6,40	7,80	8,96	8,40	9,96	11,52
6	6,33*	7,60	9,49**	9,00	10,20	13,28**
7	6,00*	7,62*	9,49**	8,85	10,37	13,28**
8	6,25	7,65	9,49**	9,00	10,35*	13,28**
9	6,22	7,81**	9,49**	8,66	11,34**	13,28**
10	6,20	7,81**	9,49**	8,60*	11,34**	13,28**
11	6,54	7,81**	9,49**	8,90*	11,34**	13,28**
12	6,16	7,81**	9,49**	8,66*	11,34**	13,28**
13	6,00	7,81**	9,49**	8,76*	11,34**	13,28**
14	6,14	7,81**	9,49**	9,00	11,34**	13,28**
15	6,40	7,81**	9,49**	8,93	11,34**	13,28**

NOTES

1 The Friedman test could be tabulated in the case of $P = 2$. However, in this case, it is sufficient to apply the binomial distribution (or its normal approximation) to the number of times that one of the two samples is preferred to the other. This is in effect the paired comparison test ("two-sided": see ISO 5495).

2 The quantity F may have only discontinuous values, with this discontinuity being very pronounced for small values of (J, P). Consequently, it is not possible to obtain critical values corresponding exactly to the levels 0,05 and 0,01; the values marked with an asterisk (*) correspond to levels slightly higher than 0,05 and 0,01; unmarked values correspond to real levels of less than 0,05 and 0,01.

3 The values marked with a double asterisk (**) are critical values obtained by means of approximation using the χ^2 distribution.

When the number P of products is greater than 5, this approximation is also used; the values of χ^2 with $(P - 1)$ degrees of freedom are given in table 4.

Table 4 — Critical values of χ^2 distribution (levels of 0,05 and 0,01)

Number of samples (or products) P	Number of degrees of freedom of χ^2 $(\nu = P - 1)$	Level of significance α	
		$\alpha = 0{,}05$	$\alpha = 0{,}01$
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	31,41	37,57
22	21	32,67	38,93
23	22	33,92	40,29
24	23	35,17	41,64
25	24	36,41	42,98
26	25	37,65	44,31
27	26	38,88	45,64
28	27	40,11	46,96
29	28	41,34	48,28
30	29	42,56	49,59
31	30	43,77	50,89

NOTE — Where, in exceptional cases, P is greater than 31, the approximate critical values are obtained using the following formulae:

— for $\alpha = 0{,}05$: $P - 0{,}15 + 1{,}645\sqrt{2P - 3}$

— for $\alpha = 0{,}01$: $P + 1{,}20 + 2{,}326\sqrt{2P - 3}$

10.3.1.2 Tied rankings

If one or more rankings are tied, F can be replaced by

$$F' = \frac{F}{1 - \{E/[JP(P^2 - 1)]\}}$$

in which E is obtained as follows.

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

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

For example, in table 2 there are two groups of tied rankings:

— 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:

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

As $J = 5$ and $P = 4$, carry out the test, having calculated F, using the value

$$F' = \frac{F}{1 - \{30/[5 \times 4(4^2 - 1)]\}} = 1{,}1\,F$$

Then compare F' with the critical values from table 3 or 4.

10.3.1.3 Comparison of two samples following the Friedman test

Where an overall difference between all the samples has been statistically demonstrated, the rank sums of each sample can be used to identify the significant differences for the sample pairs.

Let i and j be two samples, and R_i and R_j their rank sums.

Using a normal approximation, it can be decided that the two samples are different if

$$|R_i - R_j| > 1{,}960 \sqrt{\frac{JP(P+1)}{6}} \quad \text{(level 0,05)}$$

$$|R_i - R_j| > 2{,}576 \sqrt{\frac{JP(P+1)}{6}} \quad \text{(level 0,01)}$$

This test can be applied to

$$\frac{P(P-1)}{2} \text{ pairs of samples}$$

These combined tests provide a useful indication of the order in which the P samples can be hierarchically ranked, noting that the overall risk of an incorrect conclusion increases rapidly when these tests are carried out simultaneously.[1)]

10.3.2 Test of a predetermined order on the samples: Page test

It may be the case that the samples are structured by a natural order resulting, for example, from a measurable characteristic (proportions of components, temperatures, different storage times, etc.). In order to test the effect of this characteristic, it is possible to use the Page test, which is also based on rank sums, and which (in this particular case) is more powerful than the Friedman test.

If $r_1, r_2, \ldots, r_P$ are the theoretical mean ranks of the P samples placed in the predetermined order, the null hypothesis of absence of differences between the samples can be written:

$$H_0: r_1 = r_2 = \ldots = r_P$$

The alternative hypothesis is

$$H_1: r_1 \leq r_2 \leq \ldots \leq r_P$$

where at least one of these inequalities is strict.

In order to test this hypothesis, calculate

$$L = R_1 + 2R_2 + 3R_3 + \ldots + PR_P$$

and compare L with the critical values in table 5.

If L is equal to or greater than the critical value corresponding to the number of assessors, the number of samples and the chosen level of $\alpha = 0{,}05$ or $\alpha = 0{,}01$, it is concluded that the ranking drawn up by the assessors corresponds to the predetermined order of the samples.

For non-tabulated cases, calculate

$$L' = \frac{12L - 3JP(P+1)^2}{P(P+1)\sqrt{J(P-1)}}$$

This quantity approximately follows the standard normal distribution.

The alternative hypothesis is accepted if

$L' > 1{,}645$ (at the 0,05 level)

$L' > 2{,}326$ (at the 0,01 level)

11 Test report

The test report shall include the following information:

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

- number of samples;
- whether reference samples have been used;

b) the test parameters adopted:

- number of assessors and their level of qualification;
- test environment;
- material conditions;

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

d) reference to this International Standard;

e) deviations from this International Standard;

f) the name of the person supervising the test;

g) the date and time of the test.

1) This method, together with a specific application in which one of the samples is a reference sample, can be found in *Nonparametric statistical methods* (Hollander and Wolfe, 1976. Wiley).

Table 5 — Critical values from the Page test

Number of assessors J	Number of samples (or products) P											
	3	4	5	6	7	8	3	4	5	6	7	8
	Level of significance $\alpha = 0{,}05$						Level of significance $\alpha = 0{,}01$					
2	28	58	103	166	252	362	—	60	106	173	261	376
3	41	84	150	244	370	532	42	87	155	252	382	549
4	54	111	197	321	487	701	55	114	204	331	501	722
5	66	137	244	397	603	869	68	141	251	409	620	893
6	79	163	291	474	719	1 037	81	167	299	486	737	1 063
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*

NOTES

1 The Page test could be tabulated in the case of $P = 2$. However, in this case, it is sufficient to apply the binomial distribution (or its normal approximation) to the number of times that one of the two samples is preferred to the other. This is in effect the paired comparison test ("one-sided": see ISO 5495).

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

Annex A
(normative)
Practical example of application

A.1 The results of eight assessors having tested one series of samples are compiled in table A.1.

The value F from the Friedman test is calculated as follows:

as $J = 8$, $P = 5$, $R_1 = 17$, $R_2 = 31$, $R_3 = 32$, $R_4 = 23$, $R_5 = 17$,

$$F = \frac{12}{8 \times 5 \times (5 + 1)} (17^2 + 31^2 + 32^2 + 23^2 + 17^2) - $$
$$- 3 \times 8 \times (5 + 1)$$
$$= 10{,}60$$

The value 10,60 is greater than that given in table 3 for $J = 8$, $P = 5$ for a level of significance of 0,05 (i.e. 9,49); it can therefore be concluded that, with a risk of error of up to 5 %, the five samples have been recognized as being different. If a level of significance of 0,01 had been chosen, for which the critical value given in table 3 is 13,28, it would have been concluded that, with risk of error of up to 1 %, no difference had been demonstrated between the samples.

A.2 It can also be decided that two individual samples are different if the difference, as an absolute value, in their rank sums is greater than:

$$1{,}96 \times \sqrt{\frac{8 \times 5 \times (5 + 1)}{6}} = 12{,}40 \text{ (level of 0,05)}$$

or

$$2{,}576 \times \sqrt{\frac{8 \times 5 \times (5 + 1)}{6}} = 16{,}29 \text{ (level of 0,01)}$$

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

31 − 17 = 14

32 − 17 = 15

31 − 17 = 14

32 − 17 = 15

None of these sample pairs would show any significant difference if one were working with a 1 % risk of error.

This last analysis could result in the following presentation:

A E D B C

The significance of the underlining is as follows:

— two samples which are not connected by continuous underlining are different (0,05 level);

— two samples which are connected by continuous underlining are not different;

— A and E, which are not distinguished, are ranked significantly before B and C, which are themselves not distinguished. There are two groups, one containing A

Table A.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
Rank sums	17	31	32	23	17