INTERNATIONAL STANDARD

ISO 11132

First edition
2012-11-01

Sensory analysis — Methodology — Guidelines for monitoring the performance of a quantitative sensory panel

Analyse sensorielle — Méthodologie — Lignes directrices pour le contrôle de la performance d'un jury sensoriel quantitatif

Reference number
ISO 11132:2012(E)

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 11132 was prepared by Technical Committee ISO/TC 34, *Food products*, Subcommittee SC 12, *Sensory analysis*.

Sensory analysis — Methodology — Guidelines for monitoring the performance of a quantitative sensory panel

1 Scope

This International Standard gives guidelines for monitoring and assessing the overall performance of a quantitative descriptive panel and the performance of each member.

A panel of assessors can be used as an instrument to assess the magnitude of sensory attributes.

Performance is the measure of the ability of a panel or an assessor to make valid attribute assessments across the products being evaluated. It can be monitored at a given time point or tracked over time. Performance comprises the ability of a panel to detect, identify, and measure an attribute, use attributes in a similar way to other panels or assessors, discriminate between stimuli, use a scale properly, repeat their own results, and reproduce results from other panels or assessors.

The methods specified allow the consistency, repeatability, freedom from bias and ability to discriminate of panels and assessors to be monitored and assessed. Monitoring and assessment of agreement between panel members is also covered. Monitoring and assessment can be carried out in one session or over time.

Monitoring performance data enables the panel leader to improve panel and assessor performance, to identify issues and retraining needs or to identify assessors who are not performing well enough to continue participating.

The methods specified in this International Standard can be used by the panel leader to appraise continuously the performance of panels or individual assessors.

This International Standard applies to individuals or panels in training as well as for established panels.

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 8586, *Sensory analysis — General guidelines for the selection, training and monitoring of selected and expert assessors*

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

3 Terms and definitions

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

3.1
agreement
ability of different panels or assessors to assign similar scores on a given attribute to samples of the same product

3.2
homogeneity
measure of the agreement of responses among individual assessors within a test session, as a panel of assessors in replicate sessions, or for an individual assessor in replicate sessions

3.3
assessor bias
tendency of an assessor to give scores which are consistently above or below the true score when that is known or the panel mean when it is not

3.4
outlier
an assessment that does not conform to the overall pattern of the data or is extremely different from other assessments of the same or similar products

3.5
panel drift
phenomenon where a panel, over time, changes in sensitivity or becomes susceptible to biases and as a consequence changes the location on the scale where an attribute is rated for a constant, reference product

3.6
performance
ability of a panel or an assessor to make valid and reliable assessments of stimuli and stimulus attributes

3.7
repeatability
agreement in assessments of equivalent product samples under the same test conditions by the same assessor or panel

3.8
reproducibility
agreement in assessments of equivalent product samples under different test conditions, with different tasks or by a different assessor or panel

NOTE Reproducibility may be measured as any of the following:

— the reproducibility of a panel in the short term, measured between two or more sessions separated by several days;

— the reproducibility of a panel in the medium or long term, measured among sessions separated by several months;

— the reproducibility between different panels, in the same laboratory or in different laboratories;

— the reproducibility of assessments by a single assessor of different attributes of a product.

3.9
validation
process of establishing that sensory data correlate with other data on samples of the same product (e.g. laboratory measurements, consumer perception, results from other panels, consumer complaints) or that a panel or assessor is able to meet specified performance criteria

3.10
session
occasion on which products are assessed

NOTE In a single session either one or several products may be assessed by one or several assessors. For an assessor, whether alone or as part of a panel, sessions are separated in time.

3.11
replicate sessions
sessions in which the assessors, the products, the test conditions, and the task are the same

4 Principle

This International Standard is concerned with sensory panels used to assess the magnitude of one or more sensory attributes in order to make quantitative descriptions or profiles of products. Different methods are appropriate to the assessment and monitoring of the performance of panels used for difference testing.

The performance of a quantitative sensory panel may be evaluated by using assessments already available or from panel sessions conducted specifically for the purpose of obtaining performance data.

This International Standard may be used either for periodic monitoring or for reviewing ongoing profile data.

A dedicated monitoring procedure at periodic intervals is appropriate for accreditation and other purposes. Figure 1 is a flow chart for this procedure.

To review ongoing profile data generated by a panel, it can be appropriate to use data that originated from quite different profiling experiments using different product types, product numbers, etc. The procedure is the same as that shown in Figure 1. However, as there are no predefined differences, it is recommended that attributes that are significantly discriminated by the panel as a whole for a given profile be used as the key measures to check the performance of individual panelists. Attributes that result in no significant difference cannot be reliably used to check consistency since the lack of agreement within and between panelists probably means that the products are very similar for those characteristics.

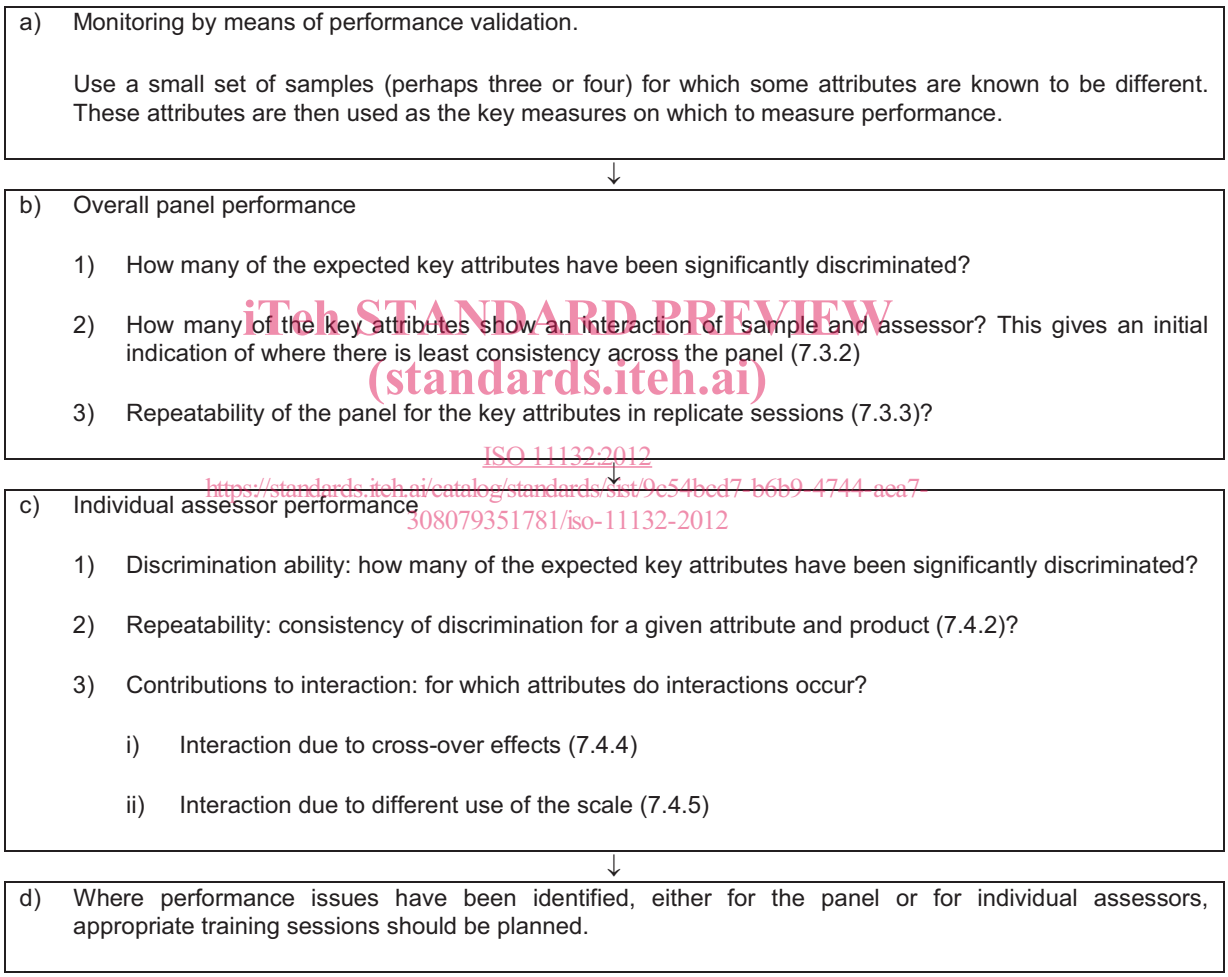

Figure 1 — Flow chart for performance monitoring

In a single session, the following indicators can be determined.

— *Bias of an assessor*, measured as the difference between the assessor's mean and a known, 'true' value, or the mean of the panel as an estimate of the 'true' value.

— *Repeatability of an assessor*, inversely related to the standard deviation (SD) of repeat assessments by the assessor of the same sample, or between replicates of the same product.

— *Reproducibility of an assessor*, inversely related to the SD of the assessor's biases across individual products.

— *Discrimination of an assessor*, measured as the ability to assign consistently different scores to different products.

Bias in an assessor may indicate sensory sensitivity that is different from other assessors and/or use of the response scale in a way that differs from other assessors.

If an assessor appears to give assessments that differ from those of other assessors, review all the results with a view to determining whether:

a) the assessments are consistent or variable for repeated samples of the same product;

b) the assessments are similar or different for samples of different products;

c) bias occurs with all, or only some, assessment scales.

Analysis of variance (ANOVA) can be used to investigate these questions.

In some cases, bias may indicate an assessor of superior ability whose results are particularly useful. In other cases, an assessor showing bias may require retraining or removal from the panel.

A single, consistent approach to statistical analysis of the results is described here. However, some attributes of panel performance can be assessed by more than one descriptive measure. For instance, error mean square and error SD (its square root) both express variability in the evaluation of a product. The measures used should be those that are usual in the field of application.

Other relevant measures of agreement between assessors in the use of the scale for an attribute are the interaction of assessor and product and the coefficient of correlation between an assessor's scores and the panel means. An assessor may have no bias, but may be using the scale in a different way. A correlation close to 1, a regression slope close to 1, and a regression intercept close to 0 indicate good agreement between an assessor and the rest of the panel.

With a small number of assessments (fewer than six) the correlation coefficient should be interpreted with caution, as it can be high (up to 0,7), by chance alone.

5 Experimental conditions

The test facilities shall be in accordance with ISO 8589.

6 Qualification of assessors

The panel shall have the level of qualification and experience of selected assessors (ISO 8586) or better.

7 Procedure

7.1 Monitoring via formal performance validation

At each session, the panel of assessors should be presented with a set of samples similar to those the panel are to assess when evaluating products and for which statistically significant differences between at least one pair of the samples can be guaranteed for at least eight attributes.

This number is recommended to encourage panel leaders or sensory managers to identify and select validation samples that show a realistic as well as a statistical measure of a panel's performance.

These key attributes are used as key measures against which to assess panel performance. The sample set should include replicates. There shall be the same number of replicates of each sample. The numbers of assessors, samples, and replicates depends on the products, the sensory attributes assessed and the purpose of the procedure. For example 2 or 3, replicates of three or four samples might be used. Care should be taken to limit the number of assessments required so as to avoid sensory fatigue. The attributes of the samples should be similar to the range of values that the panel assesses when evaluating products.

A randomized block experimental design has been adopted, in which the assessors are the “blocks”.

If there is expected to be a carry-over effect from one sample to the next, a suitable experimental design is the Williams Latin square. The basic design uses four assessors and four samples.

Table 1 — Williams Latin square

Assessor	Order			
	1	2	3	4
1	A	B	C	D
2	B	D	A	C
3	C	A	D	B
4	D	C	B	A

In this design, each assessor samples the four products in a different order and any particular product is followed by a different one for each assessor, for example A is followed by B for assessor 1, C for assessor 2, D for assessor 3 and none for assessor 4.

If multiples of four assessors are available, the same design can be repeated for each set of four.

7.2 Statistical analysis of data from formal performance validation (a single session)

Table 2 illustrates one way to tabulate and summarize the results. Some computer software may require a different organization of the data, for instance with the samples in columns and the assessors in rows.

Table 2 — Results of the assessors

Sample	Assessor								Mean
	1		2		j		n_q		
	Scores	Mean	Scores	Mean	Scores	Mean	Scores	Mean	
1	Y_{111} Y_{112} Y_{11n_r}	$\bar{Y}_{11.}$			Y_{1j1} Y_{1j2} Y_{1jn_r}	$\bar{Y}_{1j.}$			$\bar{Y}_{1..}$
2									
i	Y_{i11} Y_{i12} Y_{i1n_r}	$\bar{Y}_{i1.}$			Y_{ij1} Y_{ij2} Y_{ijn_r}	$\bar{Y}_{ij.}$			$\bar{Y}_{i..}$
n_p									
Mean					$\bar{Y}_{.j.}$				$\bar{Y}_{...}$

In this table it is assumed that there are:

$n_p \equiv$ number of samples ($i = 1,2 \ldots n_p$);

$n_q \equiv$ number of assessors ($j = 1,2 \ldots n_q$);

$n_r \equiv$ number of replicates per sample ($k = 1,2 \ldots n_r$).

Measures of the performance of the panel as a whole and individual assessors, other than bias, require the data to be analysed by ANOVA.

The details of the basic calculations are not shown in this International Standard, since the analyses are normally carried out by a computer package.

Each assessor's data are analysed by one-way ANOVA (Table 3).

Table 3 — ANOVA for an individual assessor for one attribute

Source of variation	Degrees of freedom	Sum of squares	Mean square	F-ratio
Between samples	$\nu_1 = n_p - 1$	S_1	$MS_1 = s_1/\nu_1$	$F = MS_1/MS_2$
Error	$\nu_2 = n_p(n_r - 1)$	S_2	$MS_2 = s_2/\nu_2$	
Total	$\nu_3 = n_p n_r - 1$	S_3		
$n_p \equiv$ number of samples $n_r \equiv$ number of replicates per sample				

The data for the complete session are analysed by randomized block ANOVA (Table 4).

Table 4 — ANOVA for a complete session for one attribute

Source of variation	Degrees of freedom	Sum of squares	Mean square	F-ratio
Between samples	$\nu_4 = n_p - 1$	S_4	$MS_4 = s_4/\nu_4$	
Between assessors	$\nu_5 = n_q - 1$	S_5	$MS_5 = s_5/\nu_5$	$F = MS_5/MS_7$[a]
Interaction	$\nu_6 = (n_p - 1)(n_q - 1)$	S_6	$MS_6 = s_6/\nu_6$	$F = MS_6/MS_7$
Error	$\nu_7 = n_p n_q(n_r - 1)$	S_7	$MS_7 = s_7/\nu_7$	
Total	$\nu_8 = n_p n_q n_r - 1$	S_8		
$n_p \equiv$ number of samples $n_q \equiv$ number of assessors $n_r \equiv$ number of replicates per sample				
[a]If the interaction is significant, the F-ratio for between assessors is calculated by $F = MS_5/MS_6$ with the interaction mean square in the denominator.				

7.3 Overall panel performance from formal performance validation

7.3.1 Key attribute discrimination

The proportion of key attributes that have been significantly discriminated as expected should be determined. For each attribute, this is indicated by significant variation between samples at a level of 0,05 in the ANOVA table for a session (Table 4). The higher the proportion of key attributes significantly discriminated, the better the panel is performing. The panel should receive further training on key attributes that are not significantly discriminated as expected.

7.3.2 Homogeneity of the panel

A panel is not homogeneous when any assessors are in disagreement with the rest of the panel.

A panel is not homogeneous if the interaction of sample and assessor in the ANOVA is significant at a level of 0,05.

The degree of homogeneity of the panel is inversely related to the interaction SD, s_i.

$$s_i = \sqrt{\frac{MS_6 - MS_7}{n_r}}$$

See Table 4.

The number of key attributes giving significant interaction of sample and assessor should be determined. Refer to the ANOVA table for each attribute and note those showing interaction at a level of 0,05. The higher the

number of key attributes giving significant interaction, the less consistently the panel is performing. The panel should receive further training on key attributes that are giving significant interaction.

7.3.3 Repeatability of the panel

The repeatability of the panel can be estimated from the repeatability of the individual assessors. This is inversely related to the error SD, s_e:

$$s_e = \sqrt{MS_7}$$

See Table 4.

7.3.4 Reproducibility of the panel

To check for reproducibility of the panel, make evaluations of other samples of the same products at different sessions.

The "between-sessions" factor in a three-way ANOVA (samples, assessors, sessions) should not be significant at a level of 0,05.

The interaction of samples and sessions should not be significant at a level of 0,05. If it were significant it would indicate that the evaluation of differences between samples was changing from session to session.

The interaction between assessors and sessions should not be significant at a level of 0,05. If it were significant it would indicate that the biases of individual assessors were varying from session to session.

If the analysis is being used to describe the performance of the panel as a whole, then the factors in the ANOVA (sessions, samples and assessors) are random factors. The component SDs may be combined to give a measure of reproducibility:

Reproducibility SD, s_R:

$$s_R = \sqrt{s_e^2 + s_a^2 + s_{sess}^2 + s_{a\times sess}^2 + s_{prod\times sess}^2}$$

where

e represents error;

a represents assessors;

sess represents sessions;

prod represents products.

Estimates of bias and variation can be tabulated and/or plotted. Plots over time will show if drifts, step changes or occasional problems have occurred.

Examples of such presentations are cusum analysis (see Annex B) and Shewhart control charts (see Annex C).

7.4 Individual assessor performance from formal performance validation

7.4.1 Discrimination ability of an assessor

Discrimination ability is measured by the proportion of expected key attributes that have been significantly discriminated. For each attribute, this is indicated by "between samples" variation significant at a level of 0,05 in the ANOVA table (Table 3). The higher the proportion of key attributes significantly discriminated, the better the assessor is performing. The assessor should receive further training on expected key attributes that are not significantly discriminated.