INTERNATIONAL STANDARD

ISO 11132

Second edition
2021-09

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

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

Reference number
ISO 11132:2021(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword iv
Introduction vi
1 Scope 1
2 Normative references 1
3 Terms and definitions 1
4 Principle 3
4.1 Two possible approaches 3
4.1.1 General 3
4.1.2 Performance measurement via a dedicated procedure 3
4.1.3 Ongoing monitoring via routine product profiling 4
4.2 Indicators of panel or individual assessor performance 4
4.3 Statistical analyses 5
5 Prerequisites 5
5.1 Experimental conditions 5
5.2 Qualification of assessors 5
6 Performance measurement via a dedicated procedure 5
6.1 Sample and attribute selection 5
6.2 Experimental designs 5
6.2.1 General 5
6.2.2 Randomized block design 6
6.2.3 Balanced and random designs 6
6.2.4 Same order design 6
6.3 Statistical analyses 7
6.4 Performance of the overall panel — Interpretation of statistical output 9
6.4.1 Key attribute discrimination 9
6.4.2 Agreement at panel level 9
6.4.3 Repeatability of the panel 10
6.5 Performance of individual assessors — Interpretation of statistical output 10
6.5.1 Discrimination ability of an assessor 10
6.5.2 Repeatability of an assessor 10
6.5.3 Consistency of an assessor 10
6.5.4 Agreement among assessors 11
6.5.5 Bias — Different use of scale 11
6.6 Performance issues 12
6.6.1 General 12
6.6.2 Panel 12
6.6.3 Individual assessor 12
6.7 Experimental design for following up the performance over time 12
7 Procedure for ongoing monitoring via routine product profiling 12
7.1 Attribute selection 12
7.2 Statistical analyses 12
7.3 Following up the performance over time 12
7.4 Statistical analysis of data over time 12
7.5 Statistical analysis of complete profiles 13
Annex A (informative) **Example of a practical application** 14
Bibliography 22

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.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

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. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 34, *Food products*, Subcommittee SC 12, *Sensory analysis*, in collaboration with the European Committee for Standardization (CEN) Technical Committee CEN/SS C01, *Food Products*, in accordance with the Agreement on technical cooperation between ISO and CEN (Vienna Agreement).

This second edition cancels and replaces the first edition (ISO 11132:2012), which has been technically revised. The main changes compared with the previous edition are as follows:

- the title has been changed to specify that the document is applicable to descriptive sensory panels;
- the Scope has been revised:
  - in order to provide a distinction of application for validation and monitoring, with improved wording to clarify;
  - it has been reduced to measure repeatability only, and reproducibility has been stated to be out of scope;
  - the type of quantitative descriptive sensory panels for which the document is applicable to has been specified;
- the definitions have been revised and new terminological entries have been added;
- the process for the dedicated procedure has been improved;
- experimental designs have been reviewed and augmented;
- statistical analyses related to analysis of variance have been reviewed and augmented to include more models, especially regarding sessions and panellists (fixed or random) effects and interactions;
- the subclauses (specifically the original 6.4.4 and 7.4) and Annexes B and C related to reproducibility have been removed to align with the changes to the Scope.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

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

Performance is the measure of the ability of a panel or an assessor to make reliable and valid attribute assessments across the products being evaluated. It can be assessed at a given time point, typically after a training period (validation) or tracked over time (monitoring). Performance comprises the ability of a panel to detect, identify and measure an attribute, use attributes in a similar way to other panels or between assessors within a panel, discriminate between stimuli, use a scale properly, repeat their own results, and reproduce results in comparison to other panels or assessors.

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

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

1 Scope

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

This document is applicable to the validation of the training of individual assessors or panels, as well as to the performance monitoring of established panels.

This document does not apply to the panel performance for descriptive methods where the individual scores of each assessor are not recorded, where there is no single list of attributes that is common to all the assessors, or where dominance rather than intensity is measured. Consequently, the performance of descriptive panels using methods such as consensus profile, free-choice profile, flash profile and temporal dominance of sensations (TDS) are out of scope.

The methods specified in this document are for monitoring and assessing the ability of a panel and its assessors to discriminate between products, the agreement between assessors of the same panel and the repeatability of these assessors in their intensity scoring.

Reproducibility, including both the comparison between panels and the comparison within the same panel of several evaluations conducted under different conditions (i.e. separated in time), is out of scope of this document.

The methods specified in this document can be used, in full or a selection only, by the panel leader to appraise continuously the performance of panels or individual assessors. The methods listed are not exhaustive and other appropriate methods can also be used.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements 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 document, the terms and definitions given in ISO 5492 and the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

3.1
agreement
ability of different panels or assessors to exhibit the same product differences when assigning scores on a given attribute to the same set of products

3.2
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

3.3
performance
ability of a panel or an assessor to make reliable and valid assessments of stimuli and stimulus attributes

3.4
validation
process of establishing that a panel or assessor is able to meet specified *performance* (3.3) criteria

3.5
session
period of time in which products are assessed

Note 1 to entry: 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.

[SOURCE: ISO 5492:2008/Amd.1:2016, 4.63]

3.6
replicate
occurrence of a particular condition in an experimental design

Note 1 to entry: The term usually implies that the occurrence is one of several of the same kind, but it can refer to a single occurrence. When the condition is performed twice, the wording is “two repetitions”, etc.

Note 2 to entry: To specify more than one occurrence of a condition, the terms “replication” or “replicate session” are more explicit.

Note 3 to entry: A “replicate session” is a *session* (3.5) in which the assessors, products, test conditions and task are the same.

3.7
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

[SOURCE: ISO 5492:2008/Amd.1:2016, 1.40]

3.8
order bias
arising from a product’s spatial or temporal position relative to a group of products being assessed

Note 1 to entry: The term includes both “position bias” and “sequential bias”.

[SOURCE: ISO 5492:2008/Amd.1:2016, 1.42]

3.9
repeatability
agreement (3.1) in assessments of the same products under the same test conditions by the same assessor or panel

Note 1 to entry: Repeatability can be measured within one *session* (3.5) or over several distinctly separate sessions, provided that the *replicate* (3.6) evaluations are conducted under the test conditions that can be considered to be the same. If replicate evaluations are conducted in distinctly separate sessions/sittings, the sessions are generally separated by several days only. In this case, the distinction between repeatability and *reproducibility* (3.10) in the short term is minor and relates to the test conditions being considered the same or not.

[SOURCE: ISO 5492:2008/Amd.1:2016, 1.45, modified — Note 1 to entry has been added.]

3.10
reproducibility
agreement (3.1) in assessments of the same products under different test conditions or by different assessors or panels

Note 1 to entry: Reproducibility may be measured as any of the following:

- the reproducibility of a panel (or an assessor) in the short term, measured between two or more *sessions* (3.5) separated by several days;
- the reproducibility of a panel (or an assessor) 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.

[SOURCE: ISO 5492:2008/Amd.1:2016, 1.46]

4 Principle

4.1 Two possible approaches

4.1.1 General

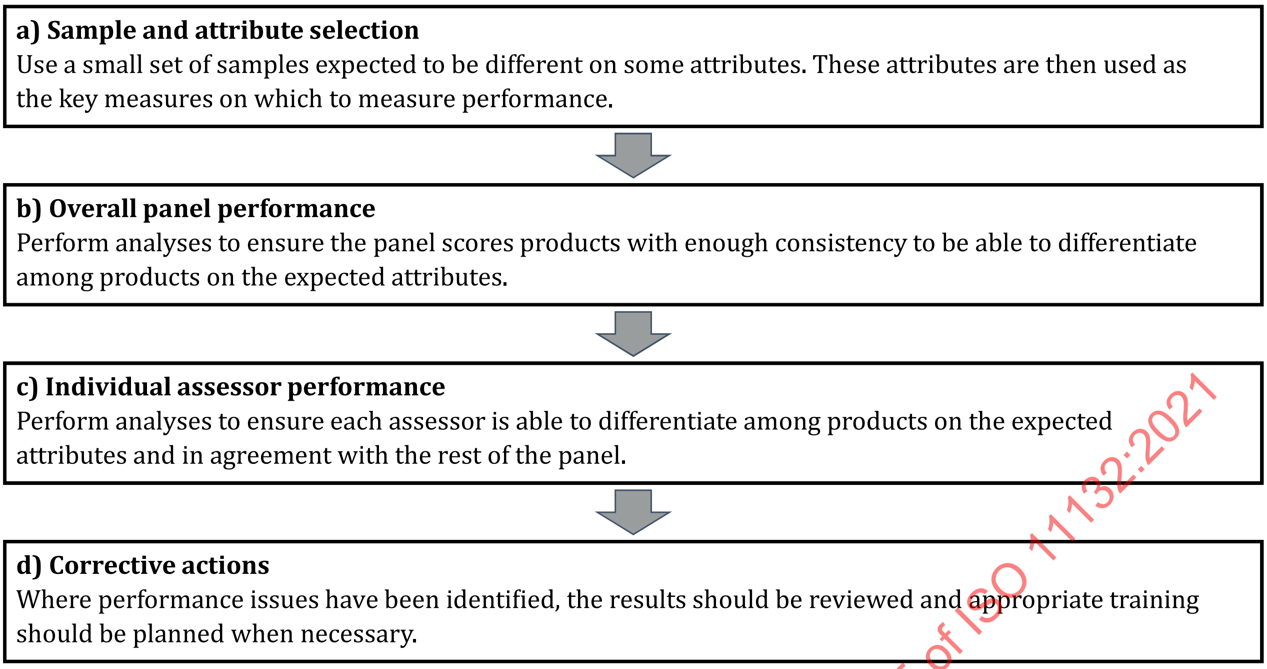
This document is concerned with sensory panels used to assess the magnitude of one or more sensory attributes in order to make quantitative descriptions or profiles of products (see ISO 13299). Different methods are appropriate to the measurement of the performance of panels used for difference testing.

The performance of a quantitative sensory panel may be evaluated from panel sessions conducted specifically for the purpose of obtaining performance (called “dedicated procedure”) by using assessments already available (called “ongoing monitoring”).

4.1.2 Performance measurement via a dedicated procedure

A dedicated procedure is the method of choice for the certification of individual assessors and for other validation purposes. For the certification renewal, this dedicated procedure should be repeated at periodic intervals, as needed. Figure 1 is a flow chart for this procedure.

This approach can typically be used at the end of the training phase of a panel to ensure that the panel and the individual assessors have achieved the desired level of performance and can be considered as trained sensory assessors or expert sensory assessors (depending on the performance criteria).



a) Sample and attribute selection

Use a small set of samples expected to be different on some attributes. These attributes are then used as the key measures on which to measure performance.

b) Overall panel performance

Perform analyses to ensure the panel scores products with enough consistency to be able to differentiate among products on the expected attributes.

c) Individual assessor performance

Perform analyses to ensure each assessor is able to differentiate among products on the expected attributes and in agreement with the rest of the panel.

d) Corrective actions

Where performance issues have been identified, the results should be reviewed and appropriate training should be planned when necessary.

Figure 1 — Process steps for the performance measurement via a dedicated procedure

4.1.3 Ongoing monitoring via routine product profiling

Another approach consists of monitoring profile data that was already collected. 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 for which the products 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 sensory assessors. Attributes that result in no significant difference cannot be reliably used to check consistency since the lack of agreement within and between sensory assessors probably means that the products are very similar for those characteristics.

In this case, over a given period, it will be necessary to check on a set of products more different that the panel is indeed capable of highlighting difference in these characteristics.

4.2 Indicators of panel or individual assessor performance

For one assessment, the following indicators can be determined:

- discrimination of the panel, measured as the ability of the panel to exhibit significant differences among products;
- discrimination of an assessor, measured as the ability of the assessor to exhibit significant differences among products;
- agreement of an assessor, measured as the degree of alignment between the assessor's average product scores and the ones of the panel;
- agreement of the panel, measured as the degree of alignment between the assessors' average product scores.

For replicate assessments:

- repeatability of an assessor, measured as the degree of homogeneity between replicated assessments of the same product;

— repeatability of a panel, measured as the average degree of homogeneity between replicated assessments of the same product for each assessor.

4.3 Statistical analyses

A single, consistent approach to statistical analysis of the results is described in this document. However, some indicators of panel performance can be assessed by more than one measure. For instance, error mean square and error standard deviation (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 use 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.

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

5 Prerequisites

5.1 Experimental conditions

The test facilities should be in accordance with ISO 8589.

5.2 Qualification of assessors

The panel should have the level of qualification and experience of selected/screened assessors in accordance with ISO 8586 or higher.

6 Performance measurement via a dedicated procedure

6.1 Sample and attribute selection

At each dedicated study, 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 are expected for each of the relevant attributes.

In order to ensure that all key aspects of the products are examined, an adequately diverse set of attributes should be included in the test.

These relevant attributes are used as key measures against which to assess panel performance. The sample set should include replicates. There should be the same number of replicates of each sample. The replicates can be evaluated within a single session or over two or more sessions. The number of assessors, samples and replicates depends on the products, sensory attributes assessed and purpose of the procedure. For example, two or three replicates of three or four samples might be used. Care should be taken to limit the number of assessments required in a session, 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.

6.2 Experimental designs

6.2.1 General

Several types of experimental designs can be used in the dedicated procedure, depending on the most important objective to answer.

6.2.2 Randomized block design

A randomized block experimental design can be used, in which the assessors are the “blocks”. This design is appropriate when no carry-over effect from one sample to the next is expected. Otherwise, a balanced design should be used instead (see 6.2.3).

6.2.3 Balanced and random designs

If a carry-over effect is expected from one sample to the next, a suitable experimental design is the Williams Latin square[12]. Table 1 shows the Williams Latin square design with four assessors and four samples.

Table 1 — Williams Latin square design for four assessors and four samples

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

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

For each replicate of the products’ evaluation, it is recommended to use a different product order for each assessor, in order to reduce the order effect and the carry-over effect.

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

It is also possible to choose a random product order design, i.e. to randomly affect each product to each position in each session.

The advantage of these approaches is to minimize the carry-over effect at panel level and therefore get better estimates of the product means at panel level for performance evaluation. However, if the product order does have an impact, the agreement between assessors will be impacted because each assessor will not experience the same product order. In order to compare the assessors on the exact same task, the same product order can be used for all sensory assessors (see 6.2.4).

6.2.4 Same order design

In order to focus on individual assessor performance and, in an effort to compare assessors under the most similar conditions, an alternative design is proposed, whereby all assessors evaluate the products in the same order, see Table 2.

Table 2 — Same order design for four assessors and four samples

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

It is worth mentioning that in this case the assessors are not evaluating the products per se but the products at a given position (product and position effects are confounded). This will lead to a biased estimate of the product effect (biased by the order effect), but to an unbiased estimate of the assessor effect and the product*assessor interaction.

6.3 Statistical analyses

Table 3 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 3 — Results of the assessors on one attribute

Sample	Assessor								Mean
	1		...	j		...	n_q		
	Scores	Mean		Scores	Mean		Scores	Mean	
1	Y_{111} Y_{112} ... Y_{11n_r}	$\bar{Y}_{11.}$		Y_{1j1} Y_{1j2} ... $Y_{1\,jn_r}$	$\bar{Y}_{1\,j.}$				$\bar{Y}_{1..}$
...									
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									$\bar{Y}_{n_p\,..}$
Mean	$\bar{Y}_{.1.}$			$\bar{Y}_{.j.}$			$\bar{Y}_{.n_q.}$		$\bar{Y}_{...}$

In this table it is assumed that:

n_p = number of samples (i = 1,2 ... n_p);

n_q = number of assessors (j = 1,2 ... n_q);

n_r = number of replicates per sample (k = 1,2 ... 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 analysis of variance (ANOVA)[7].

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

Each assessor's data are analysed by one-way ANOVA when the replicates evaluations are conducted within one single session (see Table 4).

If the replicate evaluations are conducted in distinctly separate sessions, and depending on the standard practice of the researcher/laboratory, a one-way ANOVA model (i.e. sample effect) or a two-way ANOVA model (i.e. session and sample effects) could be used (see Tables 4 and 5).

Table 4 — ANOVA for an individual assessor for one attribute

Source of variation	Degrees of freedom	Sum of squares	Mean square	F-ratio
Between samples	$v_1 = n_p - 1$	S_1	$\text{MS}_1 = s_1/v_1$	$F = \text{MS}_1/\text{MS}_2$
Error	$v_2 = n_p(n_r - 1)$	S_2	$\text{MS}_2 = s_2/v_2$	
Total	$v_3 = n_p n_r - 1$	S_3		
n_p = number of samples n_r = number of replicates per sample				

Table 5 — ANOVA for an individual assessor for one attribute with session effect

Source of variation	Degrees of freedom	Sum of squares	Mean square	F-ratio
Between samples	$v_1 = n_p - 1$	S_1	$\text{MS}_1 = s_1/v_1$	$F = \text{MS}_1/\text{MS}_5$
Between sessions	$v_4 = n_s - 1$	S_4	$\text{MS}_4 = s_1/v_1$	$F = \text{MS}_4/\text{MS}_5$
Error	$v_5 = n_p(n_r - 1) - (n_s - 1)$	S_5	$\text{MS}_5 = s_5/v_5$	
Total	$v_3 = n_p n_r - 1$	S_3		
n_p = number of samples n_r = number of replicates per sample				

The complete data set is analysed by randomized block ANOVA (see Table 6).

The assessor effect can be considered either fixed or random[8]. For performance measurement, it is usual to set the assessor effect as fixed, since the focus is on the performance of the specific assessors taking part in the study. However, in order to better predict performance under the real evaluation conditions, a random assessor factor can also be selected (see Tables 6 and 7, footnote a).

The complete data set is analysed by a two-way ANOVA when the replicate evaluations are conducted within one single session (see Table 6).

If the replicate evaluations are conducted in distinctly separate sessions/sittings, and depending on the standard practice of the researcher/laboratory, a two-way ANOVA model (i.e. panellist and sample effects) or a three-way ANOVA model (i.e. panellist, session and sample effects) can be used (see Tables 6 and 7).

An example of a practical application is given in Annex A.

Table 6 — Two-way ANOVA for a complete data set (with repetitions) for one attribute

Source of variation	Degrees of freedom	Sum of squares	Mean square	F-ratio
Between samples	$v_6 = n_p - 1$	S_6	$\mathrm{MS}_6 = s_6/v_6$	$F = \mathrm{MS}_6/\mathrm{MS}_9$[a]
Between assessors	$v_7 = n_q - 1$	S_7	$\mathrm{MS}_7 = s_7/v_7$	$F = \mathrm{MS}_7/\mathrm{MS}_9$[a]
Interaction	$v_8 = (n_p - 1)(n_q - 1)$	S_8	$\mathrm{MS}_8 = s_8/v_8$	$F = \mathrm{MS}_8/\mathrm{MS}_9$
Error	$v_9 = n_p n_q (n_r - 1)$	S_9	$\mathrm{MS}_9 = s_9/v_9$	
Total	$v_{10} = n_p n_q n_r - 1$	S_{10}		

n_p = number of samples

n_q = number of assessors

n_r = number of replicates per sample

[a] Formulae are given considering the assessor effect as fixed. Considering a random assessor effect, the denominator of the F ratio for the between samples effect becomes MS_8 instead of MS_9.

Table 7 — Three-way ANOVA for a complete data set (with repetitions) for one attribute with session effect

Source of variation	Degrees of freedom	Sum of squares	Mean square	F-ratio
Between samples	$v_6 = n_p - 1$	S_6	$\mathrm{MS}_6 = s_6/v_6$	$F = \mathrm{MS}_6/\mathrm{MS}_{12}$[a]
Between assessors	$v_7 = n_q - 1$	S_7	$\mathrm{MS}_7 = s_7/v_7$	$F = \mathrm{MS}_7/\mathrm{MS}_{12}$[a]
Between sessions	$v_{11} = n_s - 1$	S_{11}	$\mathrm{MS}_{11} = s_{11}/v_{11}$	$F = \mathrm{MS}_{11}/\mathrm{MS}_{12}$[a]
Interaction	$v_8 = (n_p - 1)(n_q - 1)$	S_8	$\mathrm{MS}_8 = s_8/v_8$	$F = \mathrm{MS}_8/\mathrm{MS}_{12}$
Error	$v_{12} = n_p n_q (n_r - 1) - (n_s - 1)$	S_{12}	$\mathrm{MS}_{12} = s_{12}/v_{12}$	
Total	$v_{10} = n_p n_q n_r - 1$	S_{10}		

n_p = number of samples

n_q = number of assessors

n_r = number of replicates per sample

n_s = number of sessions

[a] Formulae are given considering the assessor effect as fixed. Considering a random assessor effect, the denominator of the F ratio for the between samples effect becomes MS_8 instead of MS_{12}.

6.4 Performance of the overall panel — Interpretation of statistical output

6.4.1 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 an alpha level of 0,05 in the ANOVA table for a complete data set (see Tables 6 and 7). 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.

6.4.2 Agreement at panel level

A panel is not in agreement when any assessor is in disagreement with the rest of the panel (see 6.5.4).

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

The degree of agreement of the panel is inversely related to the interaction term, s_i, as shown by Formula (1):

$$s_\mathrm{i} = \sqrt{\frac{\mathrm{MS}_8 - \mathrm{MS}_9}{n_r}} \text{ or } s_\mathrm{i} = \sqrt{\frac{\mathrm{MS}_8 - \mathrm{MS}_{12}}{n_r}} \qquad (1)$$

See Tables 6 and 7.

The number of key attributes giving significant interaction of sample and assessor is determined based on the ANOVA (see Tables 6 and 7) for each attribute. The higher the number of key attributes giving significant interaction, the less consistently the panel is performing. If the interaction is significant, the nature of the interaction should be investigated at panellist level, and action taken when needed. For example, if a single panellist is disagreeing with the differences between products rated by the rest of the panel, then this panellist should be re-trained.

The nature of the interaction is often investigated by plotting the assessor by product means. Another option is to use the mixed assessor model (MAM)[6][10].

6.4.3 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 term, s_e, as shown by Formula (2):

$$s_\mathrm{e} = \sqrt{\mathrm{MS}_9} \text{ or } s_\mathrm{e} = \sqrt{\mathrm{MS}_{12}} \qquad (2)$$

depending on the chosen model (with or without a session effect).

See Tables 6 and 7.

$$s_R = \sqrt{s_\mathrm{e}^2 + s_\mathrm{a}^2 + s_\mathrm{sess}^2 + s_{\mathrm{a} \times \mathrm{sess}}^2 + s_{\mathrm{prod} \times \mathrm{sess}}^2}$$

6.5 Performance of individual assessors — Interpretation of statistical output

6.5.1 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 an alpha level of 0,05 in the ANOVA table (see Tables 4 and 5). 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.

6.5.2 Repeatability of an assessor

The repeatability of an assessor is inversely related to the assessor’s error term, s_e, as shown by Formula (3):

$$s_\mathrm{e} = \sqrt{\mathrm{MS}_2} \text{ or } s_\mathrm{e} = \sqrt{\mathrm{MS}_5} \qquad (3)$$

depending on the chosen model (with or without a session effect).

See Tables 4 and 5.

6.5.3 Consistency of an assessor

Consistency of an assessor is inversely related to the SD of the bias terms calculated from each sample.

(For assessor j, the bias term for sample i is the difference between the assessor's mean for the sample and the panel mean for the sample, $\bar{Y}_{ij.} - \bar{Y}_{i..}$. See Table 3.)

Where it is shown that an assessor's performance lacks consistency, a scatter diagram of the assessor's scores against the panel means, along with regression and correlation analysis, shows whether the inconsistency is random or has a pattern which indicates different use of the scale from the rest of the panel.

6.5.4 Agreement among assessors

A panel is not homogenous when at least one assessor is in disagreement with the rest of the panel.

This may be detected by:

- an assessor having a significant bias;
- an assessor's residual term being significantly greater than for the panel as a whole;
- the correlation coefficient between the assessor's scores and the panel means being very small or negative;
- the slope of the regression of the assessor's scores on the panel means being significantly different from 1 or the intercept being statistically significantly different from 0, or both.

Agreement among the assessors is inversely related to the between-assessors term, s_a, as shown by Formula (4):

$$s_a = \sqrt{\frac{\mathrm{MS}_8 - \mathrm{MS}_9}{n_q n_r}} \tag{4}$$

Disagreement among the assessors should be tested for significance using the "between assessors" F-ratio and comparing it with tabulated values of F for the relevant degrees of freedom. If it is significant, there is good evidence that there is a problem of panel consistency that needs to be addressed. Lack of significance does not, by itself, give reassurance that there is no problem, because it may be obscured by poor repeatability (a higher than expected error term, s_e).

6.5.5 Bias — Different use of scale

A significant ANOVA assessor bias may indicate that assessors use the scale in different ways.

In most cases, no "true" value is known and the overall bias for an assessor is taken to be the difference between that assessor's mean and the mean for the panel.

Bias for assessor *j* is given by Formula (5):

$$\bar{Y}_{.j.} - \bar{Y}_{...} \tag{5}$$

Scales (see ISO 4121) may be used by assessors in different ways. In "universal" scale use, the intensity of each attribute is rated in relation to the assessor's knowledge of the total sensory variation that can be experienced for a specific product type. Panels that work on one or only a few product categories more commonly develop this approach. In "relative" scale use, the frame of reference used by an assessor for rating intensity is related to the sensory variation shown by the set of products in a given test. This approach is more likely to be used by panels that work on a wide range of product categories. To help reduce scaling bias, it is important to ensure that the scaling approach is consistent within a panel.

6.6 Performance issues

6.6.1 General

Performance issues once identified can be listed and training sessions planned accordingly.

6.6.2 Panel

Training sessions can be organized for the panel as a whole for those attributes causing problems.

6.6.3 Individual assessor

For specific issues with individual assessor performance, it can be appropriate to discuss the problem areas privately on a one-to-one basis using a neutral or positive tone, e.g. with the provision of individual level data compared with the panel average as feedback. This can then be followed by full panel training sessions.

6.7 Experimental design for following up the performance over time

If a study is to be planned in order to evaluate the consistency of a panel over time, one session per month over a period of a year provides sufficient data. Each session should be designed with a balanced design as shown in 6.2.3. The review of the performance over time can be used to identify patterns, such as panel drift or performance improvement after re-training sessions.

7 Procedure for ongoing monitoring via routine product profiling

7.1 Attribute selection

The procedure is the same as for performance measurement via a dedicated procedure (see Clause 6). However, as there are no predefined differences, it is recommended that the attributes that are significantly discriminated by the panel as a whole for a given profile be then used as the key measures to check an individual assessor's performance. The attributes recording no significant differences cannot be reliably used to check consistency, as the lack of agreement within and between assessors is probably due to the products being very similar for those characteristics.

7.2 Statistical analyses

The statistical analyses are similar to those described for the performance measurement via a dedicated procedure (see Clause 6), except that the assessor effect should always be considered as random.

7.3 Following up the performance over time

If data from several sessions of routine assessments are already available, they can be analysed to show any changes that occurred over time. The review of the performance over time can be used to identify patterns, such as performance improvement after re-training sessions or panel drift.

7.4 Statistical analysis of data over time

The global analysis of the data over several sessions should be undertaken using repeated measures ANOVA. In practice, the same assessors may not be at all sessions, and it would be necessary to use the general linear model option of ANOVA to obtain unbiased estimates of each assessor's bias and of other parameters and components of variance.

For the panel, estimates a) and b) can be obtained.

a) Consistency of the panel can be estimated from the session term (see Table 7), if data on identical control samples has been collected over the series of sessions.

b) Internal consistency: when individual biases occur, the interaction of assessor and session measures how constant they are.

For each assessor, three estimates can be obtained in respect of each attribute.

- Overall bias: the average, over replications and/or sessions, of the differences between the assessor's scores and the corresponding means of the panel as a whole.
- Consistency: inversely related to variation of the bias terms across sessions.
- Repeatability: variation among the scores of identical samples, determined by pooling the estimates of residual SD from each session.

$$s_R = \sqrt{s_{\text{res}}^2 + s_{\text{a} \times p}^2 + s_p^2}$$

7.5 Statistical analysis of complete profiles

The methods of statistical analysis described in the preceding subclauses are applied to each attribute separately to assess the performance of the panel and the assessors on each attribute/question they had to answer.

In addition, to get an overview of the entire body of data, multidimensional techniques can be used such as principal component analysis (PCA), discriminant analysis (DA) and generalized Procrustes analysis (GPA). These approaches can be used during performance validation or ongoing monitoring. For more details on these multivariate analysis methods, see References [5], [9] and [11].

Annex A
(informative)

Example of a practical application

A.1 Data tabulation

At one session, four assessors gave scores for one attribute on three replicates of six samples. Table A.1 shows the results for this example.

NOTE This is an illustrative example. More than four assessors would normally take part.

Table A.1 — Data tabulation of the results of the assessors

Sample	Assessor								Mean
	Assessor 1		Assessor 2		Assessor 3		Assessor 4		
	Scores	Mean	Scores	Mean	Scores	Mean	Scores	Mean	
1	8 8 9	8,3	5 8 9	7,3	6 7 5	6,0	9 8 8	8,3	7,50
2	6 8 7	7,0	6 7 4	5,7	5 4 7	5,3	7 7 6	6,7	6,17
3	4 5 5	4,7	5 2 3	3,3	4 3 5	4,0	5 5 5	5,0	4,25
4	6 6 5	5,7	6 4 6	5,3	4 2 4	3,3	6 5 5	5,3	4,92
5	4 5 3	4,0	3 2 4	3,0	4 4 5	4,3	4 5 4	4,3	3,92
6	5 6 6	5,7	4 2 7	4,3	5 4 6	5,0	7 5 7	6,3	5,33
Mean	5,89		4,83		4,67		6,00		5,35

A.2 Statistical analyses

The ANOVA tables are detailed for this example in Tables A.2, A.3, A.4 and A.5.

Table A.2 — ANOVA for complete data set (assessor = fixed effect)

Source of variation	Degrees of freedom	Sum of squares	Mean square	F-ratio
Between samples	5	104,90	20,98	16,39[a]
Between assessors	3	26,04	8,68	6,79[a]
Interaction	15	16,04	1,07	0,84[b]
Residual	48	61,33	1,28	
Total	71	208,31		

[a] Significant at the level $\alpha = 0{,}05$.

[b] Not significant at the level $\alpha = 0{,}05$.

Table A.3 — Analysis of variance — Individual assessors

Source of variation	Degrees of freedom	Assessor							
		Assessor 1		Assessor 2		Assessor 3		Assessor 4	
		MS	*F*	MS	*F*	MS	*F*	MS	*F*
Between samples	5	7,42	13,36[a]	7,83	2,66[b]	2,80	2,40[b]	6,13	13,80[a]
Residual	12	0,56		2,94		1,17		0,44	
Residual SD, *s*		0,75		1,71		1,08		0,67	

[a] Significant at the level $\alpha = 0{,}05$.

[b] Not significant at the level $\alpha = 0{,}05$.

Table A.4 — Individual biases and residual SDs

Assessor	Bias	Residual SD
1	5,89 − 5,35 = +0,54	0,75
2	4,83 − 5,35 = −0,52	1,71
3	4,67 − 5,35 = −0,68	1,08
4	6,00 − 5,35 = +0,65	0,67

NOTE The bias is the difference between the assessor's mean and the overall mean, both given in Table A.1.

Table A.5 — Individual sample bias terms

Sample	Assessor			
	1	2	3	4
1	0,83	−0,17	−1,50	0,83
2	0,83	−0,50	−0,83	0,50
3	0,42	−0,92	−0,25	0,75
4	0,75	0,42	−1,58	0,42
5	0,08	−0,92	0,42	0,42
6	0,33	−1,00	−0,33	1,00
SD, *s*	0,31	0,56	0,78	0,24

NOTE An individual bias is the difference between an assessor's mean for a sample and the panel mean for that sample, both given in Table A.1.

A.3 Performance of the overall panel — Interpretation of the statistical output

From Table A.2, it can be seen the “between samples” effect is significant (at an alpha level of 0,05), indicating that the panel is able to show consistent differences among products.

From the same table, it can also be seen that the interaction was not significant at the alpha level of 0,05, indicating that the panel members were not significantly inconsistent in their differences.

The significant “between-assessors” F-ratio shows that assessors gave different scores on average (across all products). The degree of variation in assessor means can be described by the assessor SD.

In this example, the same conclusions would have been reached with a random assessor effect rather than a fixed assessor effect. Some differences of interpretation may occur when the product*assessor interaction is significant.

A.4 Performance of individual assessor — Interpretation of the statistical output

A.4.1 General

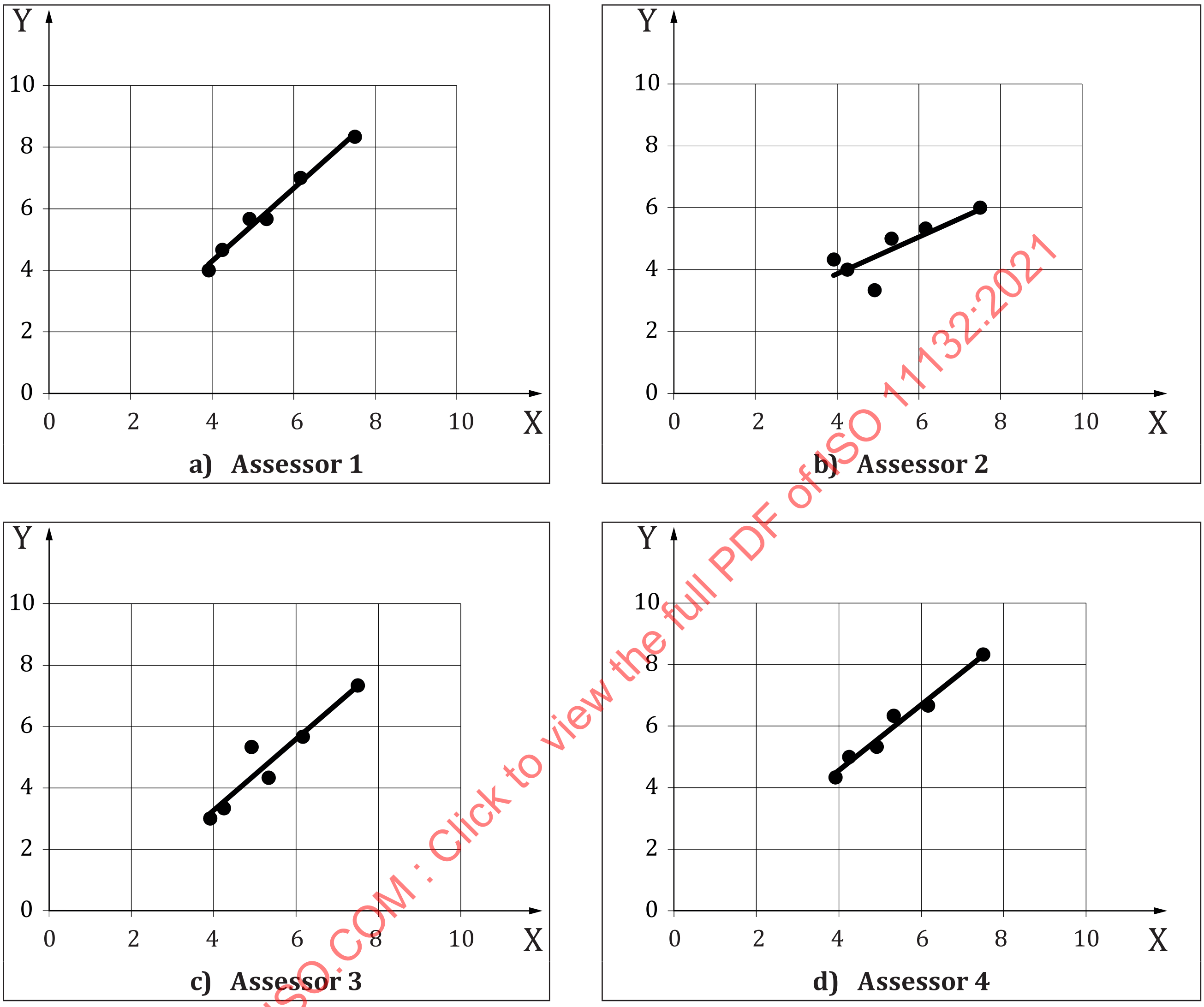
Assessors 2 and 3 had the highest residual SD (see Table A.4), indicating a lower repeatability among the replicates of the same sample than Assessors 1 and 4.

Assessor 3 also had, on average, a high negative bias, indicating a tendency to give lower scores than the rest of the panel. This assessor was also less consistent than others, varying from 1,58 below the panel mean to 0,42 above the panel mean, with the largest SD value (0,78).

Assessor 4 had a high positive bias of +0,65, but was consistent as the SD of biases was only 0,24. Since Assessors 1 and 4 agree well and have low variability, it is likely that their scores are trustworthy and the panel mean has been lowered by Assessors 2 and 3, so the “bias” of Assessor 4 is not a cause for concern.

A.4.2 Regression and correlation statistics

Figure A.1 shows each assessor's scores of the six products plotted against the panel means.



Key

X panel mean

Y assessor score

Figure A.1 — Assessor 1/2/3/4 versus panel mean

In this example, there are no "true" scores. The panel mean is used as the reference score for each assessor.

The ideal plot is one showing complete agreement between an assessor and the panel mean, with points close to a line of slope, b = 1,00, and intercept, a = 0,00. The correlation coefficient should be close to +1,00.

The regression and correlation statistics for the four assessors are shown in Table A.6.

Table A.6 — Regression and correlation statistics

Parameter	Assessor			
	1	**2**	**3**	**4**
Correlation	0,99	0,95	0,81	0,99
Slope, *b*	1,18	1,16	0,59	1,07
x-intercept, *a*	−0,42	−1,36	1,49	0,29

Assessor 4 appears to be the best, with a correlation coefficient close to 1, a slope close to 1 and the smallest intercept.

Assessor 3 had a small slope, indicating a narrower use of the scale than the other assessors.

Assessor 2 had a negative intercept, indicating a negative bias.

A.5 Additional performance issues

A.5.1 General

Line graphs may be useful to reveal problems needing further investigation.

A.5.2 Individual assessor

Three examples to compare the performance of individual assessors in a panel are shown in Figures A.2 to A.4.

Figure A.2 shows a situation where there is generally good agreement for sample separation for all but one assessor. Assessor 10 has little discrimination between samples. The remaining assessors show good agreement for all samples apart from Sample A.