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 iv

1 Scope 1

2 Normative references 1

3 Terms and definitions 1

4 Principle 2

5 Experimental conditions 4

6 Qualification of assessors 4

7 Procedure 4
7.1 Monitoring via formal performance validation 4
7.2 Statistical analysis of data from formal performance validation (a single session) 5
7.3 Overall panel performance from formal performance validation 6
7.4 Individual assessor performance from formal performance validation 7
7.5 Performance issues 9
7.6 Monitoring via routine product profiling 9
7.7 Experimental design for study of performance over time 9
7.8 Statistical analysis of data over time 9
7.9 Reproducibility between panels 10
7.10 Statistical analysis of complete profiles 10

Annex A (informative) Example of practical application 11

Annex B (informative) Example of use of cusum analysis 18

Annex C (informative) Example of use of Shewhart chart 21

Bibliography 23

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

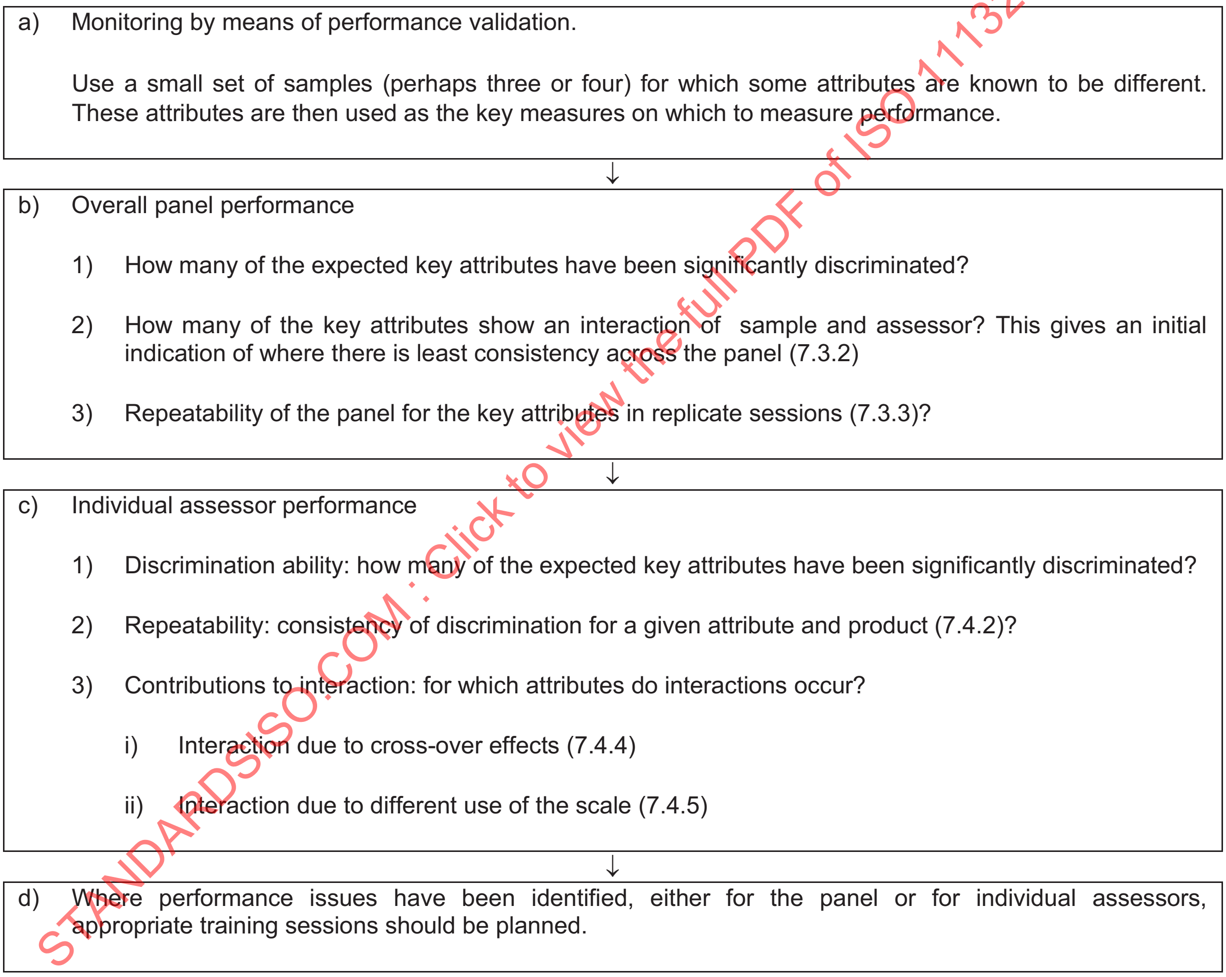
This International Standard is concerned with sensory panels used to assess the magnitude of one or more sensory attributes in order to make quantitative descriptions or profiles of products. Different methods are appropriate to the assessment and monitoring of the performance of panels used for difference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



a) Monitoring by means of performance validation.

Use a small set of samples (perhaps three or four) for which some attributes are known to be different. These attributes are then used as the key measures on which to measure performance.

↓

b) Overall panel performance

1) How many of the expected key attributes have been significantly discriminated?

2) How many of the key attributes show an interaction of sample and assessor? This gives an initial indication of where there is least consistency across the panel (7.3.2)

3) Repeatability of the panel for the key attributes in replicate sessions (7.3.3)?

↓

c) Individual assessor performance

1) Discrimination ability: how many of the expected key attributes have been significantly discriminated?

2) Repeatability: consistency of discrimination for a given attribute and product (7.4.2)?

3) Contributions to interaction: for which attributes do interactions occur?

i) Interaction due to cross-over effects (7.4.4)

ii) Interaction due to different use of the scale (7.4.5)

↓

d) Where performance issues have been identified, either for the panel or for individual assessors, appropriate training sessions should be planned.

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	$v_1 = n_p - 1$	S_1	$MS_1 = s_1/v_1$	$F = MS_1/MS_2$
Error	$v_2 = n_p(n_r - 1)$	S_2	$MS_2 = s_2/v_2$	
Total	$v_3 = n_p n_r - 1$	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	$v_4 = n_p - 1$	S_4	$MS_4 = s_4/v_4$	
Between assessors	$v_5 = n_q - 1$	S_5	$MS_5 = s_5/v_5$	$F = MS_5/MS_7$[a]
Interaction	$v_6 = (n_p - 1)(n_q - 1)$	S_6	$MS_6 = s_6/v_6$	$F = MS_6/MS_7$
Error	$v_7 = n_p n_q(n_r - 1)$	S_7	$MS_7 = s_7/v_7$	
Total	$v_8 = n_p n_q n_r - 1$	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.

$$s_\mathrm{i} = \sqrt{\frac{MS_6 - MS_7}{n_r}}$$

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:

$$s_\mathrm{e} = \sqrt{MS_7}$$

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:

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

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.

7.4.2 Repeatability of an assessor

The repeatability of an assessor is inversely related to the assessor's error SD, s_e:

$$s_\mathrm{e} = \sqrt{MS_2}$$

See Table 3.

7.4.3 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,) $\overline{Y}_{ij.} - \overline{Y}_{i..}$. See Table 2.

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.

7.4.4 Agreement among assessors

A panel is not homogenous when one or more assessors is in disagreement with the rest of the panel.

This may be detected by:

- an assessor having a significant bias (see Annex B);
- an assessor's residual SD 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 and/or the intercept being statistically significantly different from 0.

Agreement among the assessors is inversely related to the between-assessors SD, s_a.

$$s_\mathrm{a} = \sqrt{\frac{MS_5 - MS_7}{n_q n_r}}$$

if the interaction was not significant (see Table 4) or

$$s_\mathrm{a} = \sqrt{\frac{MS_5 - MS_6}{n_q n_r}}$$

if the interaction was significant. See Table 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 SD, s_e).

7.4.5 Different use of scale/bias

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:

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

Scales (see ISO 4121[2]) 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.

7.5 Performance issues

7.5.1 General

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

7.5.2 Panel

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

7.5.3 Individual assessor

For specific issues with individual assessor performance, it may be appropriate to discuss the problem areas privately on a one-to-one basis first and follow through with full panel training sessions.

7.6 Monitoring via routine product profiling

The procedure is the same as for monitoring via formal performance validations (see 7.1 to 7.5). However, as there are no predefined differences, it is recommended that the attributes which are significantly discriminated by the panel as a whole for a given profile be then used as the key measures to check individual panellists’ performance. The attributes recording no significant difference 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.

7.7 Experimental design for study of 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 as in 7.1.

If data from several sessions of routine assessments are already available, they can be analysed to show any changes that occurred over time.

7.8 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-to-session SD, 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, estimates 1) to 3) can be obtained in respect of each attribute.

1) 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.

2) Consistency — inversely related to variation of the bias terms across sessions.

3) Repeatability — variation among the scores of identical samples, determined by pooling the estimates of residual SD from each session.

7.9 Reproducibility between panels

This aspect arises only when the same products are assessed by two or more panels in separate sessions.

The statistical analysis for one attribute would be three-factor ANOVA (product, session, and panel) with a nested effect of assessors within panel.

A measure of the reproducibility between panels is the reproducibility SD, s_R:

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

where

res represents residual;

a represents assessors;

p represents panels.

7.10 Statistical analysis of complete profiles

The methods of statistical analysis described in the preceding are applied to each attribute separately. This has the benefit that assessors having problems in evaluating particular attributes can be identified.

Also, a better understanding of the entire body of data can be achieved by considering all the data summaries (measures and plots) together, using such statistical methods to analyse the complete profiles as principal component analysis (PCA), discriminant analysis (DA) and generalized Procrustes analysis (GPA).

The discriminating ability of a panel can be shown from PCA by the number of principal components in which the "between products" interaction is significant in two-factor ANOVA. The higher the number, the better the discriminating ability of the panel.

Discrimination between products is also shown directly by DA.

GPA shows whether assessors have the same interpretation of all the attributes, how different their interpretations are, and how much disagreement there is between an individual assessor and the rest of the panel.

Annex A
(informative)

Example of practical application

A.1 Monitoring via formal performance validation

At one session, four assessors gave scores for one attribute on three replicates of six samples.

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

A.2 Statistical analysis

Table A.1 — Results of the assessors

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

Table A.2 — ANOVA for complete session

Source of variation	Degrees of freedom	Sum of squares	Mean square	F-ratio
Between samples	5	104,90	20,98	
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 $\alpha = 0{,}001$

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

Table A.3 — Analysis of variance — Individual assessors

Source of variation	Degrees of freedom	Assessor							
		1		2		3		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 $\alpha = 0{,}001$

[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 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 in Table A.1.

A.3 Overall panel performance

From Table A.2, it can be seen that the interaction was not significant at the 0,05 level, indicating that the panel members were consistent in their differences.

The significant “between-assessors” F-ratio in Table A.2 shows that assessors gave different scores on average. The degree of variation in assessor means can be described by the assessor SD:

$$s_{\text{a}} = \sqrt{\frac{8{,}68 - 1{,}28}{6 \times 3}} = 0{,}64$$

A.4 Individual assessor performance

A.4.1 General

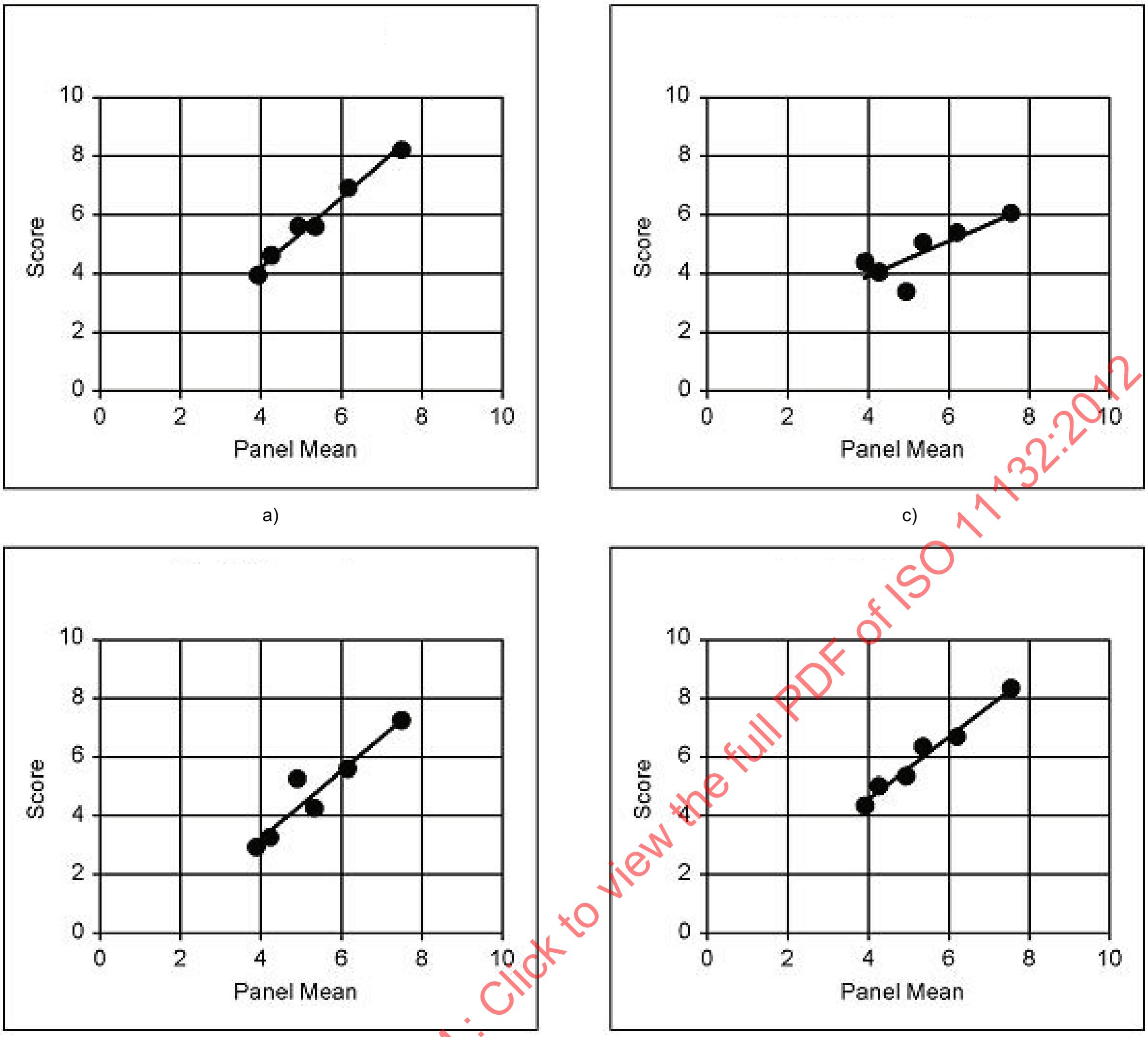
Assessors 2 and 3 had the highest residual SD (see Table A.4), indicating poor repeatability among the replicates of the same sample.

Assessor 3 also had, on average, a high negative bias, indicating a tendency to give scores lower than the rest of the panel. This assessor also was inconsistent, varying from 1,58 below the panel mean to 0,42 above the panel mean, with an SD of biases of 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 plotted against the panel means.



a)

c)

b)

d)

Figure A.1 — Scores of assessors 1 to 4 [a) to d)] plotted against the panel means

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
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 other assessors.

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

A.5 Performance issues

A.5.1 General

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

A.5.2 Panel

Two examples to compare the performance of different panels are shown. In the figures, “panels” may be different panels making assessments simultaneously or the same panel making assessments over time.

Figure A.2 shows a situation where there is generally good agreement for sample separation, but one panel with different scale usage. Panel 3 (solid triangle data points) gives, on average, lower scores than the other panels.

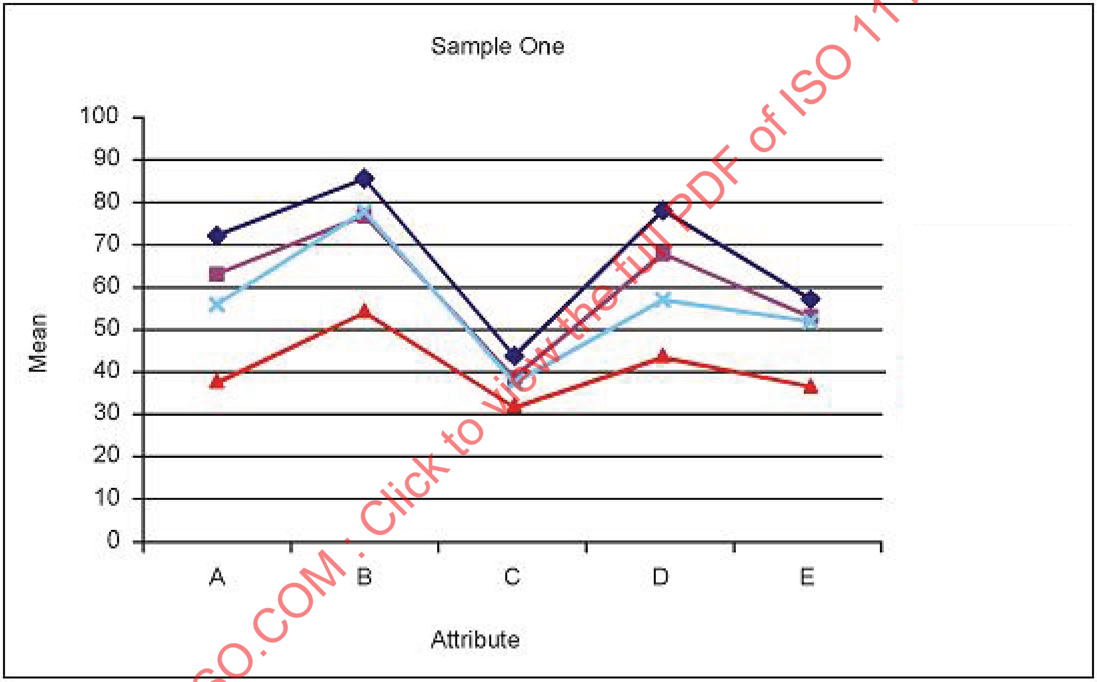


Figure A.2 — Mean scores from four panels scoring the same sample (sample 1) for five attributes

Figure A.3 shows a situation where there is poor agreement between panels for both sample separation and scale usage. Panel 2 (solid square data points) is particularly erratic in its scoring of attributes C and D in comparison with the other panels.

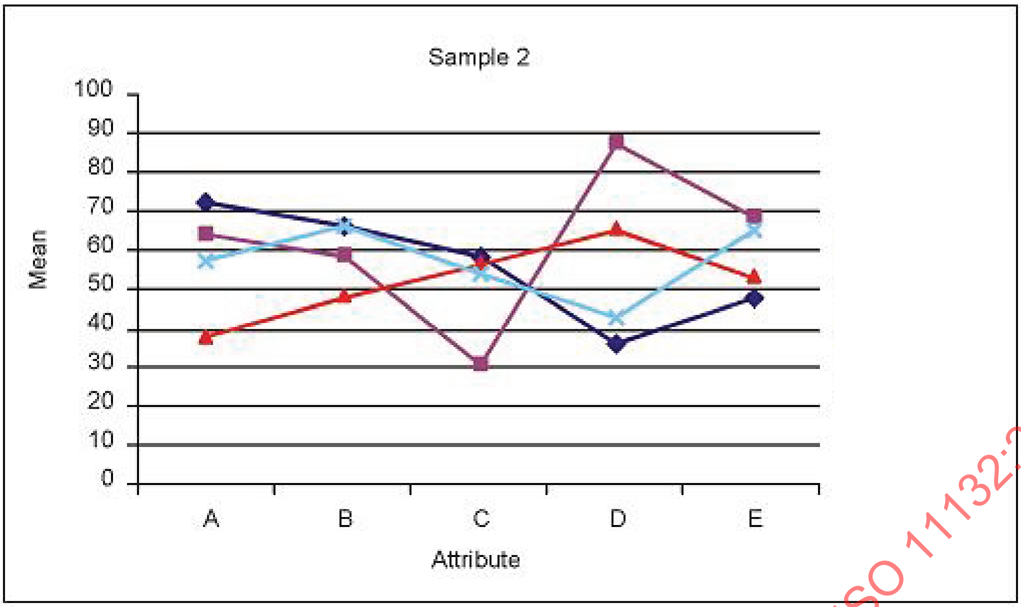


Figure A.3 — Mean scores from four panels scoring the same sample (sample 2) for five attributes

A.5.3 Individual assessor

Three examples to compare the performance of individual assessors in a panel are shown.

Figure A.4 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.

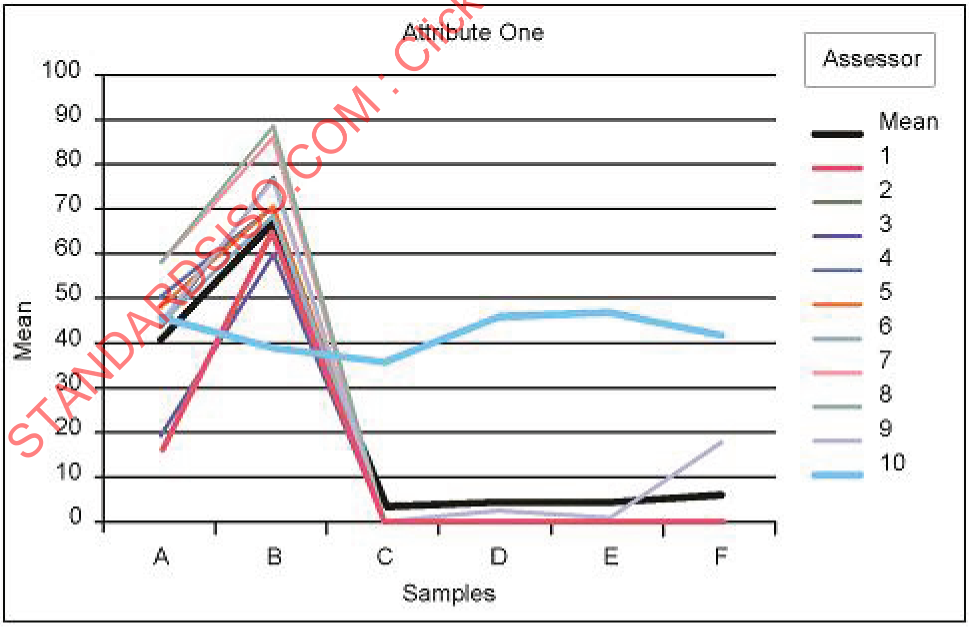


Figure A.4 — Panel scores for 10 assessors scoring six samples on one attribute (attribute 1)

Figure A.5 shows a situation where most assessors agree on the order of the samples, but assessor 10 has poor discrimination and uses little of the scale.

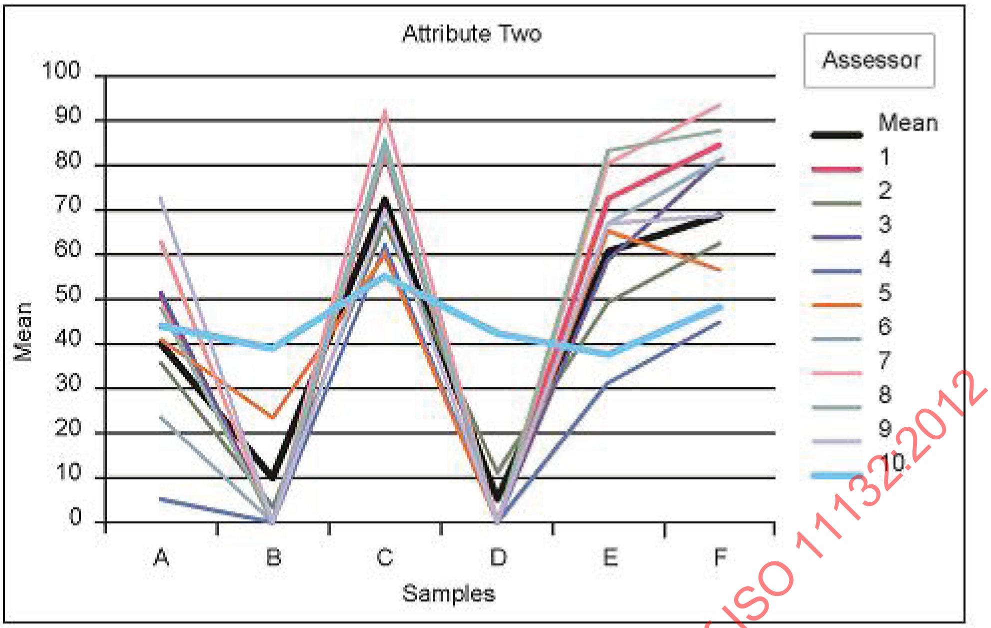


Figure A.5 — Panel scores for 10 assessors scoring six samples on one attribute (attribute 2)

Figure A.6 shows a situation where there is poor performance in both sample separation and scale usage by all assessors. The assessors show no agreement even in the ranking of the samples, and two of the assessors give very low scores to all samples.

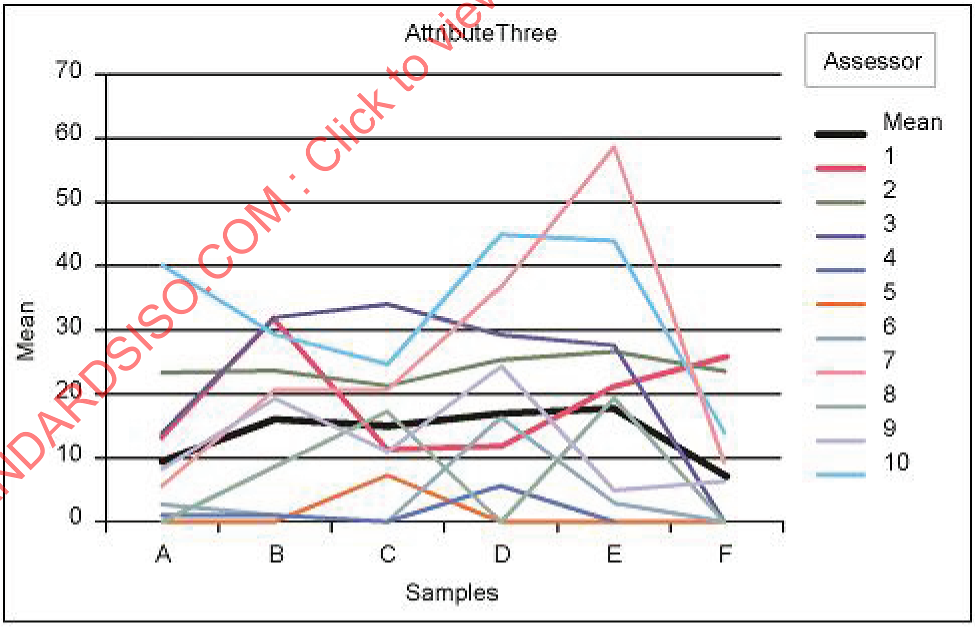


Figure A.6 — Panel scores for 10 assessors scoring six samples on one attribute (attribute 3)

Annex B
(informative)

Example of use of cusum analysis

The performance of a panel was monitored at regular intervals over a period of time and the bias of each assessor determined.

Cusum analysis was used to identify whether the performance of that assessor changed.

A "cusum" after an observation is the *cu*mulative *sum*s of differences from a target. If monitoring an assessor's bias, the target value is zero.

A change in the slope of the cusum plot corresponds to a change in the assessor's bias.

Table B.1 — Bias terms and cusums

Session	Bias	Cusum	Session	Bias	Cusum
		0,0			
1	–1,0	–1,0	16	–0,1	–1,9
2	0,0	–1,0	17	–1,0	–2,9
3	0,4	–0,6	18	–0,2	–3,1
4	0,2	–0,4	19	–0,3	–3,4
5	–1,0	–1,4	20	–0,8	–4,2
6	–0,2	–1,6	21	–1,3	–5,5
7	0,8	–0,8	22	–0,6	–6,1
8	–0,1	–0,9	23	–0,6	–6,7
9	–0,5	–1,4	24	–0,7	–7,4
10	–0,1	–1,5	25	0,4	–7,0
11	0,5	–1,0	26	0,5	–6,5
12	–0,6	–1,6	27	–0,1	–6,6
13	0,2	–1,4	28	–0,3	–6,9
14	–0,3	–1,7	29	0,0	–6,9
15	–0,1	–1,8	30	–0,5	–7,4

From the plot of the biases, it is difficult to detect any changes in the mean level, but changes in the slope of the cusum plot are much easier to detect. The means between the changes are summarized in the Manhattan diagram.