TECHNICAL REPORT

ISO/TR 14468

First edition
2010-12-15

Selected illustrations of attribute agreement analysis

Illustrations choisies d'une analyse d'accord d'attribut

Reference number
ISO/TR 14468:2010(E)

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

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

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.

In exceptional circumstances, when a technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example), it may decide by a simple majority vote of its participating members to publish a Technical Report. A Technical Report is entirely informative in nature and does not have to be reviewed until the data it provides are considered to be no longer valid or useful.

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/TR 14468 was prepared by Technical Committee ISO/TC 69, *Applications of statistical methods*, Subcommittee SC 7, *Applications of statistical and related techniques for the implementation of Six Sigma*.

Introduction

The Six Sigma[1] and statistical International Standards communities share a philosophy of continuous improvement and many analytical tools. The statistical International Standards community arrives at rigorous documents through long-term consensus. The disparities in time pressures, mathematical rigour, and statistical software usage have inhibited exchanges, synergy, and mutual appreciation between the two groups.

This Technical Report takes one specific statistical tool, attribute agreement analysis, develops the topic somewhat generically (in the spirit of International Standards), then illustrates it through the use of five detailed and distinct applications. The generic description focuses on the commonalities across studies designed to assess the agreement of attribute measurements. The annexes, containing five illustrations, follow the basic framework, but also identify the nuances and peculiarities in the specific applications.

[1] Six Sigma is a trademark of Motorola, Inc.

Selected illustrations of attribute agreement analysis

1 Scope

This Technical Report assesses a measurement process where the characteristic(s) being measured is (are) in the form of attribute data (including nominal and ordinal data).

This Technical Report provides examples of attribute agreement analysis (AAA) and derives various results to assess closeness of agreement amongst appraisers, such as agreement within appraisers, agreement between appraisers, agreement of each appraiser vs. a standard, and agreement of all appraisers vs. a standard.

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

ISO 3534-2, *Statistics — Vocabulary and symbols — Part 2: Applied statistics*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 3534-1, ISO 3534-2, and the following apply.

3.1
measurement system
collection of operations, procedures, devices and other equipment, software, and personnel used to assign a value to the characteristic being measured

[IWA 1:2005[4], 3.1.9]

NOTE In the context of this Technical Report, the personnel refer to the appraiser.

3.2
nominal data
categorical variables that have two or more levels with no natural ordering

3.3
ordinal data
categorical variables that have three or more levels with a natural ordering

3.4
binary data
categorical variables that have two levels with no natural ordering

3.5
agreement within appraiser
extent to which each appraiser agrees with himself or herself on all trials when each appraiser conducts more than one trial

3.6
agreement between appraisers
extent to which all appraisers agree with each other on all trials when more than one appraiser makes one or more appraisals

3.7
agreement of each appraiser vs. standard
extent to which each appraiser agrees with himself or herself as well as with the standard when a known standard is specified

3.8
agreement of all appraisers vs. standard
extent to which all appraisers agree with each other on all trials as well as with the standard when a known standard is specified

3.9
percentage of agreement
P %
agreement, expressed as a percentage, for multiple appraisals by one appraiser or among different appraisers

3.10
kappa
κ
statistic indicating the degree of agreement of the nominal or ordinal assessments made by multiple appraisers when evaluating the same samples

NOTE Kappa statistics are commonly used in cross-tabulation (table) applications and in attribute agreement analysis.

3.11
Fleiss's kappa
statistic used for assessing the reliability of agreement when appraiser(s) are selected at random from a group of available appraisers

3.12
Cohen's kappa
statistic used for assessing the reliability of agreement when the appraiser(s) are specifically chosen and are fixed

3.13
***p*-value**
probability of observing the observed test statistic value or any other value at least as unfavourable to the null hypothesis

[ISO 3534-1:2006, 1.49]

NOTE This concept is used in hypothesis tests to help in deciding whether to reject or fail to reject a null hypothesis.

3.14
***Z*-statistic**
test statistic which follows the standard normal distribution

4 Symbols and abbreviated terms

95 % CI	95 % confidence Interval
AAA	attribute agreement analysis
MSA	measurement system analysis
σ_κ	standard error (SE) of kappa statistic
n	sample size
P %	percentage of agreement
Z	value of the Z-statistic

5 Generic description of attribute agreement analysis

5.1 Overview of the structure of attribute agreement analysis

This Technical Report provides general guidelines on the design, conduct and analysis of studies aiming at evaluating the agreement amongst appraisers when classifying an item into two or more categories (e.g. “good” or “bad”). It describes a procedure with five steps and illustrates the steps with five distinct applications given in Annexes A to E.

The steps given in Table 1 are generic and apply to design and analysis of AAA studies in general. Each of the five steps as well as general agreement analysis methodology are explained in general in 5.2 to 5.7. Specific explanations of the substance of these steps are provided in the examples in Annexes A to E.

Table 1 — Basic steps in attribute agreement analysis

1	State the overall objectives
2	Describe the measurement process
3	Design the sampling plan
4	Analyse the result
5	Provide a conclusion with suggestions

5.2 Overall objectives of attribute agreement analysis

AAA is often used in Six Sigma projects and quality improvement projects. The primary motivation for AAA studies should be clearly stated and agreed upon by all parties. The main purpose of AAA is to evaluate the capability of a measurement system based on attribute data and to judge whether it is acceptable in the context of making correct decisions within a given monitored process. AAA determines how good agreement is among appraisers, and between appraisers and a given recognized “standard”.

AAA is conducted for a variety of reasons, which include, but are not limited to:

a) a lack of consistency in the assessment of a part or unit determined by one appraiser during different trials;

b) a lack of consistency in the assessment of a part or unit determined by different appraisers;

c) the measurement results of a part or unit determined by an appraiser or appraisers exhibiting disagreement with a known standard value for that part or unit;

d) a requirement of quality management standards, e.g. ISO/TS 16949[5].

5.3 Measurement process description

This Technical Report focuses on processes where the characteristic(s) being measured consist(s) of attribute data.

The measurement process should be clearly described before conducting AAA, including appraisers, procedures, the quality characteristic(s) to be measured, measurement conditions, and attribute data type (i.e. nominal, ordinal or binary).

5.4 Agreement analysis methodology

Many measurement processes in industry rely on gauges, weighing instruments, micrometers or other devices that make fairly direct physical measurements of a product characteristic. There are, however, many situations in which quality characteristics are difficult to define and assess, e.g. automobile performance ratings, classification of fabric quality as “good” or “bad”, and ratings of wine colour, aroma and taste on a 1 to 10 scale.

In cases when physical measurements are not possible, subjective classifications or ratings are made by people. In these situations, an AAA is needed where more than one appraiser gives a rating and an evaluation of the agreement between appraisers is made. If the appraisers agree, the possibility exists that the ratings are accurate. If the appraisers disagree, rating usefulness is limited.

The assigned ratings can be nominal, ordinal or binary. Nominal data are categorical variables that have two or more levels with no natural ordering. For example, the levels in a food tasting study may include crunchy, mushy, and crispy. Ordinal data are categorical variables that have three or more levels with a natural ordering, such as strongly disagree, disagree, neutral, agree, and strongly agree. However, distances between categories are unknown. Binary data are categorical variables that only have two levels. For instance, appraisers classify items as “good/bad”, or “go/no go”. It should be noted that binary data actually constitute a special case of nominal data with only two levels. Binary data are widely used in industry and when a standard exists giving the correct value of the unit being measured, misclassification rates can also be employed to assess the performance of a measurement system. A binary measurement system is discussed further in Annex A. Thus, in this Technical Report, nominal data refer to a variable that has three or more possible levels.

No matter what the data type is, percentage of assessment agreement can be utilized to evaluate the agreement of an attribute measurement system. Percentage of agreement quantifies the agreement for multiple ratings within one appraiser or among different appraisers. The percentage of assessment agreement, P %, is actually the point estimate for a population proportion, and is given by

$$P\ \% = \frac{n_{\text{match}}}{n} \times 100\ \%$$

where

n_{match} is the number of agreements among multiple ratings;

n is the number of samples.

For nominal data, the kappa statistic, κ, is most appropriate. It is defined as the proportion of agreement between appraisals after agreement by chance has been removed.

$$\kappa = \frac{P_{\text{obs}} - P_{\text{exp}}}{1 - P_{\text{exp}}}$$

where

P_{obs} is the observed proportion of agreement;

P_{exp} is the expected proportion due to chance agreement.

The value of kappa ranges from −1 to +1. Generally speaking, the higher the value of kappa, the stronger the agreement. If kappa has the value 1, the ratings show perfect agreement (consistency). If kappa is 0, the agreement of the ratings is the same as that expected by chance. In general, kappa values above 0,9 are considered excellent.

Kappa values less than 0,7 indicate that the rating system needs improvement, whereas those less than 0,4 indicate the measurement system capability is inadequate. Typically a kappa value of at least 0,7 is required.

The two most popular kappa statistics are Cohen's kappa, based on the two-way contingency table, and Fleiss's kappa, based on matched pairs. They treat the selection of appraisers differently when calculating the probability of agreement by chance. Cohen's kappa assumes that the appraisers are specifically chosen and are fixed, whereas Fleiss's kappa assumes that the appraisers are selected at random from a group of available appraisers. This leads to two different methods of estimating the probability. Thus kappa, and its standard error (SE), σ_κ, can be calculated with either Fleiss's method or Cohen's method. The test statistic for kappa is

$$Z = \frac{\kappa}{\sigma_\kappa}$$

with the null hypothesis $H_0{:}\kappa = 0$ and the alternative hypothesis $H_1{:}\kappa > 0$.

This is a one-sided test. Under the null hypothesis, Z follows the standard normal distribution. Reject the null hypothesis if the p-value is less than the prespecified value, commonly taken to be 0,05.

Since binary data are a special case of nominal data with only two levels, kappa statistics can also be employed to deal with a binary measurement system.

Kappa statistics do not take into account the magnitude of differences observed in ordinal data. They represent absolute agreement among ratings. Therefore, when examining ordinal data, Kendall's coefficients are the best choice. Two types of Kendall's coefficients are mentioned in this Technical Report, Kendall's coefficient of concordance (also known as Kendall's W) and Kendall's correlation coefficient (also called Kendall’s tau). Both of these coefficients are non-parametric statistics. The former, ranging from 0 to 1, expresses the degree of association among multiple ratings, whereas the latter, ranging from −1 to 1, expresses the degree of association between the known standard and a single rating. Thus, Kendall's coefficient of concordance should be used to evaluate the consistency within appraisers and between appraisers. Furthermore, when the true standard is known, Kendall's correlation coefficient can be employed to assess the following two types of agreements: agreement of each appraiser vs. standard and agreement of all appraisers vs. standard.

5.5 Sampling plan for attribute agreement analysis

In the sampling plan for AAA studies, the subgroup size of parts, the number of appraisers, and the number of trials should be determined. Generally speaking, three to five appraisers are selected to rate more than 20 parts (for multiple attributes, more samples are required to cover all the attributes) with two or three trials. Note that the selected samples should represent the entire production process. For nominal data, the appraiser selection method also determines which kappa statistic should be calculated. If the appraisers are specifically chosen and are fixed, Cohen's kappa is more appropriate. If appraisers are selected at random from a group of available appraisers, Fleiss's kappa is preferred. It is also worth mentioning that Cohen's kappa is based on the two-way contingency table. When the standard is not known, Cohen's kappa can only be calculated if and only if the data satisfy the conditions:

a) within appraiser — there are exactly two trials with an appraiser;

b) between appraisers — there are exactly two appraisers each having one trial.

In the process of measurement for AAA, randomization is a very important consideration. Randomization means the parts should be measured by the appraiser in a random order.

Table 2 shows a basic layout of an AAA with three appraisers, three repetitions, and 20 items measured by each appraiser.

Table 2 — Layout of a generic attribute agreement analysis design

Item number	Appraiser A			Appraiser B			Appraiser C			Standard
	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3	
1										
2										
3										
20										

5.6 Data analysis

The following four types of agreement need be taken into consideration:

a) within appraisers, which means that each appraiser agrees with himself or herself on all trials;

b) between appraisers, which means that all appraisers agree with each other on all trials;

c) each appraiser vs. standard, which means that each appraiser agrees with himself or herself as well as with the standard;

d) all appraisers vs. standard, which means that all appraisers agree with each other on all trials as well as with the standard.

It is quite obvious that the type of agreement c) is no less than the first one a) since it adds a constraint, namely, agreeing with the standard. The condition is quite similar for the fourth and the second types of agreements. Obviously, the fourth kind of agreement is the smallest of the four. And for each type of agreement, two types of kappa statistics are generally adopted, those of Cohen and Fleiss. Also, for nominal data with three or more categories, two types of kappa coefficients can be calculated. First, one can compute an overall kappa, which is an assessment of raters' agreement across all categories. Second, one can compute individual kappa values for each category. This reveals the categories in which raters have trouble agreeing.

In addition to the AAA report, AAA graphics are also useful. They can be used to reflect the agreement clearly and directly. Generally, the percentages of assessment agreement within and between appraisers, kappa coefficient tables, and Kendall's coefficient (ordinal data only) tables are calculated. Moreover, a graph of the matched proportions for each appraiser can be displayed when the number of trials for each appraiser is more than one. Additionally, another graph of the matched proportions between the ratings of each appraiser and the attribute can be displayed only when the attribute is known and provided for each sample.

5.7 Conclusions and suggestions

Based on the results of the AAA, a judgement can be made about the adequacy of the attribute measurement process. Generally the disagreement within an appraiser shows the appraiser cannot make consistent measurement results (possibly because the appraiser did not follow the measurement procedure exactly at different trials). The disagreement between appraisers means the appraisers' procedures are not exactly the same or the appraisers' capabilities of measurement are different (possibly due to their different experiences or physical reasons, e.g. eyesight for visual inspection). Actions shall be taken after the root cause(s) is (are) found for the inadequate attribute measurement process.

After certain actions have been taken to improve the measurement system, e.g. effective training has been done for the operators, the AAA needs to be repeated to validate whether the improved measurement system is acceptable.

6 Description of Annexes A to E

Five distinct examples of AAA are illustrated in Annexes A to E, which have been summarized in Table 3 with the different aspects indicated.

Table 3 — Example summaries listed by annex

Annex	Example	AAA details
A	LCD manufacture	Three appraisers, randomly selected among the group of appraisers, judged LCD quality on 20 samples twice by visual inspection. The inspection results are binary. Minitab[a] software package was used to perform the analysis
B	Technical support triage of issues	Nominal response with 6 categories encountered in Service Sector; 4 appraisers, no repetition, 48 issues evaluated by each appraiser. SAS JMP[b] software package was used to perform the analysis. “Truth” on correct categorization of issue is known
C	Tasting differences in water	Nominal response with 4 categories; 3 testers, 3 repetitions, leading to 12 cups of water evaluated by each tester. SAS JMP[b] software package was used to perform the analysis. “Truth” on correct categorization of brand of water is known
D	Thermistor defects	Three appraisers, randomly selected among the group of appraisers, judged 20 thermistor samples twice by visual inspection. The inspection results are nominal data, falling into 8 categories and without natural ordering. Minitab[a] software package was used to perform the analysis
E	Assessment of level of disability following a stroke	Ordinal response with 5 ordered categories encountered in the medical sector; 2 appraisers, no repetition, 46 cases evaluated by each appraiser. SAS JMP[b] software package was used to perform the analysis. “Truth” on correct categorization of issue is known

[a] Minitab is the trade name of a product supplied by Minitab, Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product named.

[b] SAS JMP is the trade name of a product supplied by the SAS Institute, Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product named.

Annex A
(informative)

Liquid crystal display manufacture

A.1 General

In a liquid crystal display (LCD) manufacturer, the display feature is judged by operators through visual inspection. All the samples are tested under video graphics array (VGA) mode. The results can be either normal colour (marked as good) or deflected colour (bad). In the measurement phase, visual inspection, leading to subjective classification, is mainly employed by the appraisers to judge whether a sample is good or bad. Therefore, the experience of the appraisers and the training they have been given are of huge importance. The objective of this study is to evaluate the consistency and accuracy of the attribute measurement system.

A.2 Response variable

The response variable is binary data (two levels with no natural ordering).

A.3 Standard attribute

Standard attribute (the correct rating) is given in this case.

A.4 Possible reasons for incorrect judgement

Failing to follow work instructions could lead to incorrect judgement. Another factor could be the experience of the appraisers and the training they have been given.

A.5 Sampling plan

To assess the consistency and accuracy of ratings, three appraisers, Carol, Fiona, and Kaka, judged LCD quality on 20 samples (model: LCD40b66) twice by visual inspection. LCD samples were randomly presented to the three appraisers, who were randomly selected from a group with the same introductory training and similar experience.

The inspection results are binary.

A.6 Raw data

Table A.1 lists the raw data used in the AAA.

Table A.1 — Inspection results of LCD and standard attribute

Part	Standard	Carol		Fiona		Kaka	
		1st Trial	2nd Trial	1st Trial	2nd Trial	1st Trial	2nd Trial
1	Good	Good	Good	Good	Good	Good	Good
2	Good	Good	Good	Good	Good	Good	Good
3	Good	Good	Good	Good	Good	Good	Good
4	Bad	Bad	Bad	Bad	Bad	Bad	Bad
5	Bad	Bad	Bad	Good	Good	Bad	Bad
6	Bad	Bad	Bad	Bad	Bad	Bad	Bad
7	Good	Good	Good	Good	Good	Good	Good
8	Good	Good	Good	Good	Good	Good	Good
9	Good	Good	Good	Good	Good	Good	Good
10	Good	Good	Good	Good	Good	Good	Good
11	Good	Good	Good	Good	Good	Good	Good
12	Good	Good	Good	Good	Good	Good	Good
13	Good	Good	Good	Good	Good	Good	Good
14	Good	Good	Good	Good	Good	Bad	Bad
15	Good	Good	Good	Good	Good	Good	Good
16	Good	Good	Good	Good	Good	Good	Good
17	Good	Good	Good	Good	Good	Good	Good
18	Good	Good	Good	Good	Good	Good	Good
19	Bad	Bad	Bad	Bad	Bad	Bad	Bad
20	Bad	Bad	Bad	Bad	Bad	Bad	Bad

A.7 Attribute agreement analysis

AAA in Minitab 15[2] is adopted to assess the consistency and accuracy of subjective classifications by examining the results within appraisers, between appraisers, and against the standard. AAA output consists of session window and graph window results.

The session window includes the following types of agreement:

a) within appraiser: it shows the consistency with which an appraiser rates the same sample across different trials;

b) between appraisers: it shows whether appraisers' ratings agree with each other, i.e. whether different appraisers give the same rating to the same sample.

Since the standard attribute (the correct rating) is given in this case, the session window output includes two additional types of agreement:

c) each appraiser vs. standard: it shows how well each appraiser's assessment of each sample matches with the standard, in other words, whether each rating of the same appraiser agrees with the standard rating;

d) all appraisers vs. standard: it shows how well responses of all appraisers agree with the known standard when they are combined.

For each type of agreement, the session window output includes assessment agreement and Fleiss's kappa statistics to assess the consistency and accuracy of the appraisers' responses.

[2] Minitab is the trade name of a product supplied by Minitab, Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product named.

A.7.1 Agreement within appraisers

The within appraisers table in the session window assists in answering whether each appraiser rated the LCD consistently across trials.

As shown in Table A.2, each appraiser rated 20 LCDs (number inspected). Carol, Fiona and Kaka evaluated 20 out of 20 LCDs the same across trials (number matched), for 100 % matched. And the 95 % confidence interval (CI) for percentage matched is 86,09 % to 100 %.

Table A.2 — Percentages of the assessment agreement within appraisers

Appraiser	Number inspected[a]	Number matched[b]	Percentage	95 % CI
Carol	20	20	100,00	(86,09, 100,00)
Fiona	20	20	100,00	(86,09, 100,00)
Kaka	20	20	100,00	(86,09, 100,00)

[a] Number of LCDs which have been rated.

[b] Number of times appraisers agree with themselves across all trials.

To evaluate the consistency of each appraiser's ratings across trials, the kappa statistic can be used within appraisers.

There are two main types of kappa statistic: Cohen's kappa is based on the two-way contingency table, while Fleiss's kappa is based on matched pairs. The two approaches treat the selection of appraisers differently when calculating the probability of agreement by chance. Cohen's kappa assumes that the appraiser(s) are specifically chosen and are fixed, whereas Fleiss's kappa assumes that the appraiser(s) are selected at random from a group of available appraisers. This leads to two different methods of estimating the probability. In this case, three appraisers were randomly selected from the whole group, thus it is not appropriate to employ Cohen's kappa to assess agreement. In the following, only Fleiss's kappa is considered.

Generally speaking, the higher the value of kappa, the stronger the agreement within appraisers. If $\kappa = 1$, this indicates perfect agreement (consistency). If $\kappa = -1$, this indicates perfect disagreement. If $\kappa = 0$, the agreement of the ratings is the same as that expected by chance. In general, kappa values above 0,9 are considered excellent. Kappa values less than 0,7 indicate that the rating system (or the service quality) needs improvement and those less than 0,4 indicate the measurement system capability is inadequate. Typically a kappa value of at least 0,70 is required, but kappa values close to 0,90 are preferred.

The p-value represents the probability of making a Type I error, which is rejecting the null hypothesis ($\kappa = 0$, or agreement within appraiser is due to chance) when the null hypothesis is true. If the p-value of a test statistic is less than the prespecified significance level (alpha), for which a commonly used value is 0,05, the null hypothesis should be rejected. Because the p-values for the three overall values of Fleiss's kappa are less than 0,05, the choice to reject the null hypothesis has to be made. The response agreements are significantly different from those expected by chance. The p-values for specific categories and appraisers are also shown in Table A.3.

Table A.3 — Fleiss's kappa statistics within appraisers

Appraiser	Response	Kappa	SE kappa	Z	p-Value (vs. >0)
Carol	Bad	1	0,223 607	4,472 14	0,000 0
	Good	1	0,223 607	4,472 14	0,000 0
Fiona	Bad	1	0,223 607	4,472 14	0,000 0
	Good	1	0,223 607	4,472 14	0,000 0
Kaka	Bad	1	0,223 607	4,472 14	0,000 0
	Good	1	0,223 607	4,472 14	0,000 0

A.7.2 Agreement of each appraiser vs. standard

It is necessary to determine how well each appraiser's assessment of each sample matches with the standard, in other words, whether each rating of the same appraiser agrees with the standard rating (see Table A.4).

Table A.4 — Assessment agreement and disagreement of each appraiser vs. standard

Assessment agreement				
Appraiser	**Number inspected**	**Number matched**[a]	**Percentage**	**95 % CI**
Carol	20	20	100,00	(86,09, 100,00)
Fiona	20	19	95,00	(75,13, 99,87)
Kaka	20	19	95,00	(75,13, 99,87)

Assessment disagreement						
Appraiser	**Number good/bad**[b]	**Percentage**	**Number bad/good**[c]	**Percentage**	**Number mixed**[d]	**Percentage**
Carol	0	0,00	0	0,00	0	0,00
Fiona	1	20,00	0	0,00	0	0,00
Kaka	0	0,00	1	6,67	0	0,00

[a] Number of times that an appraiser's assessment across trials agrees with the known standard.

[b] Ratio of good assessments across trials to bad standard assessments.

[c] Ratio of bad assessments across trials to good standard assessments.

[d] Number of non-identical assessments across trials.

The results in Table A.5 show that kappa for each appraiser is greater than 0,7, indicating that each appraiser's assessment matches well with the standard.

Table A.5 — Fleiss's kappa statistics (each appraiser vs. standard)

Appraiser	**Response**	**Kappa**	**SE kappa**	Z	p**-Value** (vs. >0)
Carol	Bad	1,000 00	0,158 114	6,324 56	0,000 0
	Good	1,000 00	0,158 114	6,324 56	0,000 0
Fiona	Bad	0,856 63	0,158 114	5,417 81	0,000 0
	Good	0,856 63	0,158 114	5,417 81	0,000 0
Kaka	Bad	0,874 61	0,158 114	5,531 51	0,000 0
	Good	0,874 61	0,158 114	5,531 51	0,000 0

A.7.3 Agreement between appraisers

The results are listed in Table A.6. The kappa value of 0,858 in Table A.7 indicates that the agreement between appraisers is acceptable.

Table A.6 — Percentages of the assessment agreement between appraisers

Number inspected	**Number matched**[a]	**Percentage**	**95 % CI**
20	18	90,00	(68,30, 98,77)

[a] Number of times that all appraisers' assessments agree with each other.

Table A.7 — Fleiss's kappa statistics between appraisers

Response	Kappa	SE kappa	Z	p-Value (vs. >0)
Bad	0,857 778	0,057 735 0	14,857 1	0,000 0
Good	0,857 778	0,057 735 0	14,857 1	0,000 0

A.7.4 Agreement of all appraisers vs. standard

The results in Tables A.8 and A.9 show the agreement with standard when the assessments of all appraisers are combined. The results indicate good match with the standard.

Table A.8 — Percentages of the assessment agreement of all appraisers vs. standard

Number inspected	Number matched[a]	Percentage	95 % CI
20	18	90,00	(68,30, 98,77)

[a] Number of times that all appraisers' assessments agree with the known standard.

Table A.9 — Fleiss's kappa statistics of all appraisers vs. standard

Response	Kappa	SE kappa	Z	p-Value (vs. >0)
Bad	0,910 413	0,091 287 1	9,973 07	0,000 0
Good	0,910 413	0,091 287 1	9,973 07	0,000 0

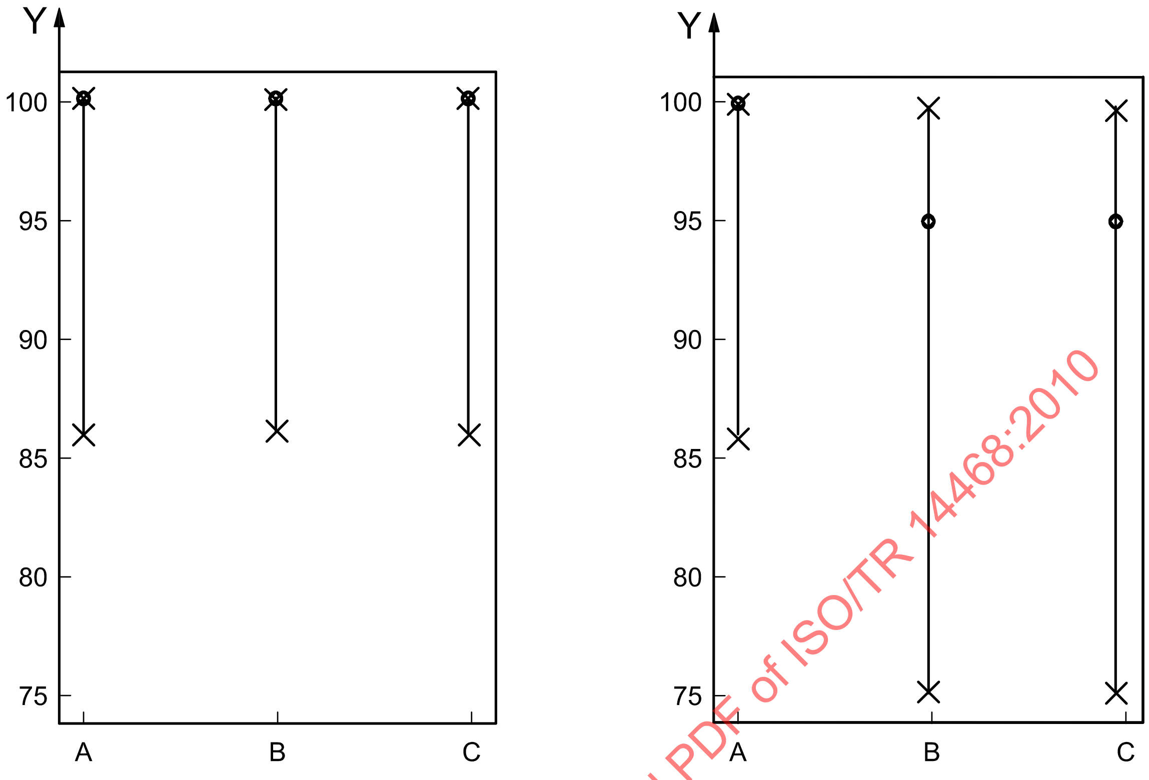
A.7.5 Figures of agreement assessment

The graph window also outputs two graphs: percentages of the assessment agreement and 95 % CI within appraisers on the left-hand side, percentages of the assessment agreement and 95 % CI of all appraisers vs. standard on the right (see Figure A.1).

Figure A.1 a) shows the consistency of each appraiser's ratings, while Figure A.1 b) also shows consistency and accuracy. The filled circles indicate the percentage matched and the lines joining the data points indicate a 95,0 % confidence interval.

A.8 Conclusions

Since all of the Fleiss's kappa statistics are greater than 0,7, the attribute data measurement system is acceptable. For continuous improvement, the standards should be reviewed with the appraisers, Fiona and Kaka. It may be necessary to provide them more training.



a) Within appraiser

b) Appraiser vs. standard

Key

Y percentage of agreement

A appraiser Carol

B appraiser Fiona

C appraiser Kaka

Figure A.1 — Percentage agreement of the assessment and 95 % CI

Annex B
(informative)

Technical support centre triage of issues

B.1 General

A software technical support centre wanted to test the ability of their first-line call receivers to correctly diagnose customer software issues into several broad categories.

At this company when a customer has an issue and places a call to technical support, the customer initially speaks with a first-line call receiver. The first-line call receivers are trained to ask questions and listen to customer answers to correctly categorize the issue into several broad categories. Once the first-line call receiver categorizes a customer issue, the call is forwarded to a technical support expert for that particular category of issues. If the first-line call receiver incorrectly categorizes an issue, the current category expert can attempt to diagnose the problem or send the caller back to a first-line call receiver. Having to re-categorize a call wastes time and can frustrate the customer. It is standard practice to record these phone calls for study and training purposes.

The objectives of the study were to determine how well the first-line call receivers categorized customer issues and to identify any areas where the first-line call receivers might need more training to determine categories by evaluating the answers to a set of standard questions.

B.2 Description of the experiment

The experiment involved tracking the categorizations of four randomly selected first-line call receivers. These first-line call receivers all had the same level of introductory training and had worked in the call centre for a period of 1 year to 1,5 years. Each of the four first-line call receivers listened to 48 different recordings of customer phone calls with various issues. They were not told how the calls were categorized at the end of the conversations. The four first-line call receivers listened to the phone call recordings and tried to categorize the problems based on the answers to the standard questions. The true category for each of these phone calls was known to the experimenters.

B.3 Response variable

The response variable is the category of the customer issue. It is a nominal response with six levels. The six levels are activation, calculations, data display, graphics, spreadsheet, and windows.

B.4 Standard attribute

The standard attribute is known and refers to the actual customer issue category for each phone call.

B.5 Measurement method

Each first-line call receiver was allowed to listen to each recorded phone call all the way through one time. They were not allowed to replay any part of the phone call. Once they had listened to the call, they had 30 s to choose an issue category.

B.6 Possible reasons for incorrect judgement

The first-line call receivers may not have had enough training on how to categorize certain types of phone calls.

B.7 Sampling plan

Four first-line call receivers (Debbie, Mark, Barbara, and Jim) were randomly selected from a group of first-line call receivers with the same introductory training and similar experience (between 1 year and 1,5 years) as the testers for the experiment.

The experimenters explained to the testers that each tester would be listening to 48 recorded customer issue phone calls. After listening carefully to each phone call they were asked to categorize the phone call within 30 s of it ending. No recorded phone calls were repeated. There were eight recorded phone calls for each issue category.

The tester categorization results and the true customer issue category are shown in Table B.1.

B.8 Raw data

The results from the experiment are shown in Table B.1.

Table B.1 — Results from the experiment

Call	Actual	Debbie	Mark	Barbara	Jim
1	Activation	Windows	Activation	Activation	Activation
2	Activation	Activation	Windows	Activation	Activation
3	Graphics	Graphics	Windows	Windows	Graphics
4	Calculations	Calculations	Calculations	Calculations	Calculations
5	Calculations	Calculations	Calculations	Calculations	Calculations
6	Windows	Windows	Windows	Windows	Windows
7	Spreadsheet	Spreadsheet	Spreadsheet	Spreadsheet	Spreadsheet
8	Data display	Spreadsheet	Data display	Data display	Calculations
9	Data display	Data display	Data display	Data display	Data display
10	Graphics	Graphics	Graphics	Graphics	Graphics
11	Spreadsheet	Spreadsheet	Spreadsheet	Spreadsheet	Spreadsheet
12	Activation	Activation	Activation	Activation	Activation
13	Graphics	Graphics	Graphics	Graphics	Graphics
14	Activation	Activation	Activation	Activation	Activation
15	Spreadsheet	Spreadsheet	Spreadsheet	Spreadsheet	Spreadsheet
16	Graphics	Graphics	Calculations	Graphics	Graphics
17	Spreadsheet	Spreadsheet	Spreadsheet	Spreadsheet	Spreadsheet
18	Activation	Activation	Activation	Windows	Windows
19	Graphics	Graphics	Windows	Graphics	Graphics
20	Windows	Graphics	Windows	Windows	Windows
21	Activation	Activation	Activation	Activation	Activation

Table B.1 (*continued*)

Call	Actual	Debbie	Mark	Barbara	Jim
22	Data display	Windows	Data display	Data display	Data display
23	Windows	Activation	Windows	Windows	Windows
24	Data display	Data display	Data display	Data display	Data display
25	Activation	Activation	Activation	Activation	Activation
26	Graphics	Graphics	Graphics	Graphics	Graphics
27	Windows	Windows	Windows	Windows	Windows
28	Calculations	Spreadsheet	Calculations	Calculations	Calculations
29	Data display	Data display	Data display	Data display	Calculations
30	Windows	Windows	Windows	Windows	Windows
31	Spreadsheet	Spreadsheet	Spreadsheet	Spreadsheet	Spreadsheet
32	Windows	Windows	Windows	Windows	Windows
33	Spreadsheet	Data display	Spreadsheet	Spreadsheet	Spreadsheet
34	Windows	Windows	Windows	Windows	Windows
35	Data display	Spreadsheet	Spreadsheet	Spreadsheet	Spreadsheet
36	Calculations	Data display	Calculations	Calculations	Calculations
37	Calculations	Calculations	Calculations	Calculations	Data display
38	Windows	Windows	Windows	Windows	Windows
39	Spreadsheet	Spreadsheet	Calculations	Calculations	Spreadsheet
40	Spreadsheet	Spreadsheet	Spreadsheet	Data display	Spreadsheet
41	Graphics	Graphics	Graphics	Graphics	Graphics
42	Data display	Spreadsheet	Data display	Data display	Data display
43	Calculations	Calculations	Calculations	Calculations	Calculations
44	Activation	Activation	Activation	Activation	Activation
45	Calculations	Data display	Data display	Calculations	Calculations
46	Calculations	Data display	Calculations	Calculations	Calculations
47	Data display	Data display	Data display	Data display	Data display
48	Graphics	Graphics	Graphics	Graphics	Graphics

B.9 Attribute agreement analysis — Within testers

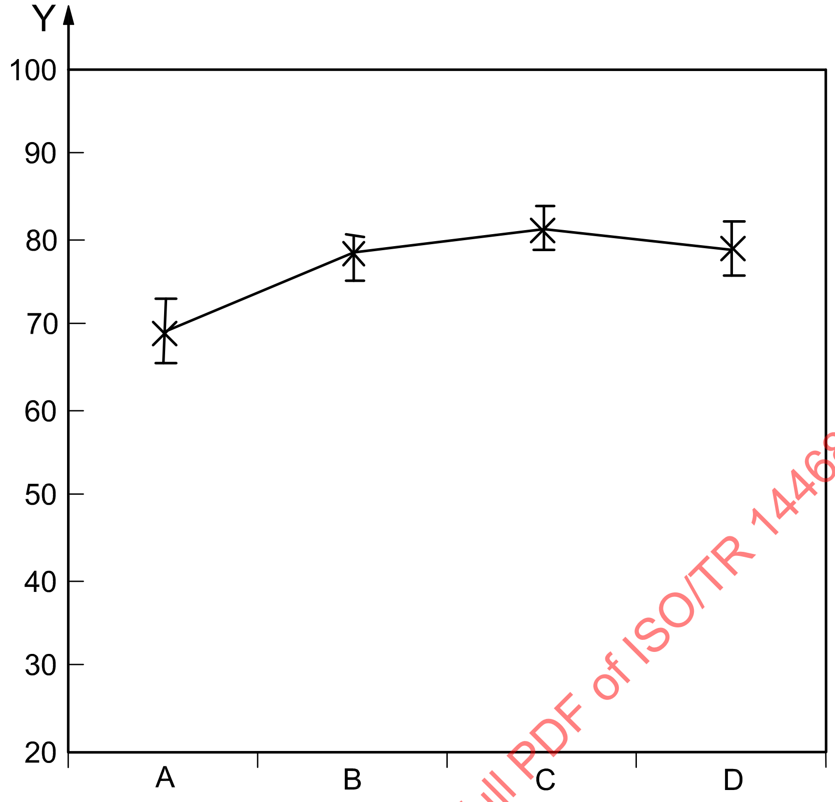
B.9.1 General

AAA in SAS JMP 9[3] is adopted to assess the consistency and accuracy of subjective classifications by examining the results within testers, between testers, and against the standard.

B.9.2 Consistency of each tester across trials

Figure B.1 and Table B.2 provide information on the percentage agreement for each tester.

[3] SAS JMP is the trade name of a product supplied by the SAS Institute, Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product named.



Key

Y percentage of agreement

A Debbie

B Mark

C Barbara

D Jim

Figure B.1 — Percentage agreement for each tester

Table B.2 — Percentage agreement for each tester with confidence intervals

Rater	Percentage agreement	95 % Lower CI	95 % Upper CI
Debbie	69,444 4	66,554 9	72,188 7
Mark	78,472 2	76,138 2	80,636 0
Barbara	81,250 0	79,135 5	83,195 7
Jim	79,166 7	76,884 0	81,278 8

B.9.3 Effectiveness of each tester across trials (against standard)

The ability to effectively categorize a customer issue is shown in Table B.3.

Table B.3 — Effectiveness of each tester

Rater	Effectiveness	95 % Lower CI	95 % Upper CI
Debbie	75,000 0	61,215 6	85,079 4
Mark	85,416 7	72,832 8	92,751 8
Barbara	89,583 3	77,832 6	95,467 8
Jim	89,583 3	77,832 6	95,467 8

The kappa statistics in Table B.4 compare each tester with the standard. Note that all testers seem to do at least an adequate job of identifying the phone call issue categories correctly. Debbie appears to struggle the most, however.

Table B.4 — Kappa statistics of each tester against standard

Rater	Compared with standard	Kappa	Standard error
Debbie	Truth	0,700 0	0,074 4
Mark	Truth	0,825 0	0,060 8
Barbara	Truth	0,875 0	0,052 8
Jim	Truth	0,875 0	0,052 7

B.10 Attribute agreement analysis — Between testers

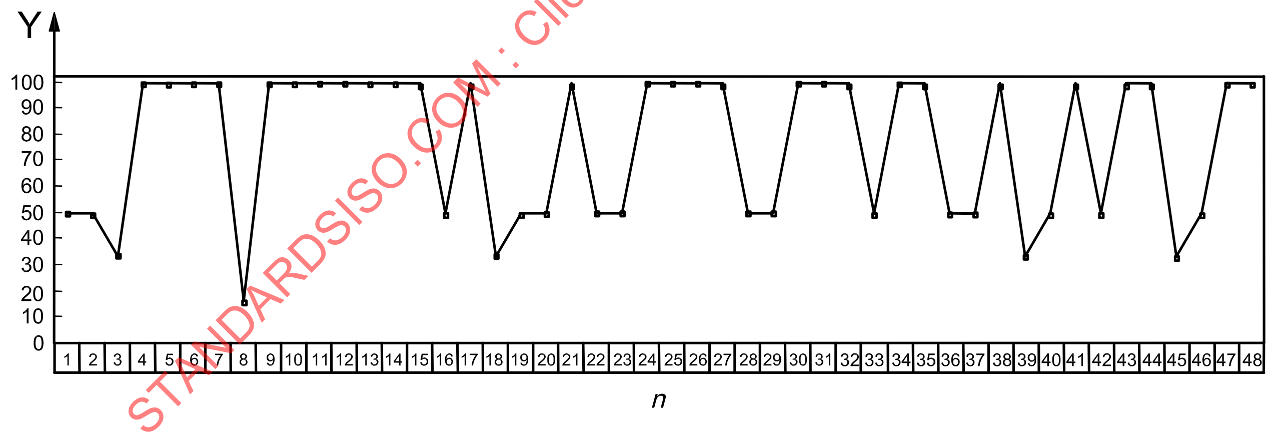
B.10.1 Consistency across testers (not against standard)

Table B.5 indicates that around 58 % of the phone calls were correctly categorized by all testers.

Table B.5 — Agreement across testers

Number inspected	Number matched	Percentage agreement	95 % Lower CI	95 % Upper CI
48	28	58,333	44,281	71,150

Figure B.2 shows the percentage agreement between testers for each recorded phone call.



Key

Y percentage of agreement

n call number

Figure B.2 — Percentage agreement between testers

Table B.6 shows the kappa statistics between each pair of testers. Debbie shows the least agreement with the other testers. The other testers have considerably better agreement with each other.

Table B.6 — Kappa statistics of each tester against another tester

Rater	Compared with rater	Kappa	Standard error
Debbie	Mark	0,626 4	0,078 8
Debbie	Barbara	0,626 6	0,079 0
Debbie	Jim	0,650 2	0,077 2
Mark	Barbara	0,849 1	0,057 5
Mark	Jim	0,749 3	0,070 2
Barbara	Jim	0,849 8	0,057 2

Table B.7 shows the kappa statistics for agreement across the issue categories. Activation and graphics issues have the most agreement. Issues about data display have the least agreement.

Table B.7 — Kappa statistics across categories

Category	Kappa	Error
Activition	0,824 0	0,058 9
Calculations	0,628 1	0,058 9
Data display	0,578 6	0,058 9
Graphics	0,824 0	0,058 9
Spreadsheet	0,778 7	0,058 9
Windows	0,715 7	0,058 9
Overall	0,724 4	0,026 5

B.10.2 Accuracy of testers against standard

Table B.8 shows the counts of category misclassifications across testers. This shows that data display issues are frequently being categorized as spreadsheet issues. Calculations issues are frequently being categorized as data display issues. Also, it appears that activation issues are sometimes misclassified as a windows issue.

Table B.8 — Misclassifications

Standard level	Activation	Calculations	Data display	Graphics	Spreadsheet	Windows
Activation	—	0	0	0	0	1
Calculations	0	—	2	1	2	0
Data display	0	5	—	0	2	0
Graphics	0	0	0	—	0	1
Spreadsheet	0	1	6	0	—	0
Windows	4	0	1	3	0	—
Other	0	0	0	0	0	0

B.11 Conclusions

The AAA study indicates that perhaps some targeted training could help improve the first-line call receivers' ability to categorize customer issues. It appears that some training to help them to differentiate between calculation, data display, and spreadsheet issues would be helpful. It also might be helpful to train the first-line call receivers in distinguishing between activation and windows issues. Overall, it seems that the first-line call receivers are doing an adequate job of categorizing the customer issues from incoming calls.

Annex C
(informative)

Tasting differences in water[4)]

C.1 General

A training organization uses the ability of people to recognize the taste of different brands of bottled water as a means to demonstrate the benefits of an agreement attribute analysis (AAA).

The objective of the study consists in having students:

a) recognize when an AAA is needed;

b) design an experiment that can be executed in class;

c) collect the data;

d) analyse and interpret the results;

e) make recommendations based on the information extracted from the analysis.

The following question was posed to the students: Can people detect differences among specific brands of bottled water as well as between these brands and tap water? To answer such a question, it is necessary to first assess the ability of people to recognize water from similar sources.

C.2 Description of the experiment

The experiment consists of selecting three different brands of bottled water (referred to as brand A, brand B, and brand C) and filling unmarked cups with water from the bottles of water or from the tap. Then “testers” blindly drink water from these cups and record the type of water (brand A, brand B, brand C, or tap water) they believed they tasted.

The purpose of the analysis is to quantify the ability of the testers to recognize consistently and effectively the various brands of bottled water or tap water.

C.3 Response variable

The response variable is the type of water tasted by a “tester”. This type of data has a nominal type with four levels in our experiment (brand A, brand B, brand C, and tap water) and no specific order.

C.4 Standard attribute

The standard attribute is known and refers to the actual type of water used to fill the cups in the experiment.

[4)] This example was supplied by Quality Management Systems Solutions (QMSS). This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of this supplier.

C.5 Measurement method

All testers were given marked samples of each of one of the four types of water at the beginning of the experiment, so they were “trained” in recognizing each type of water. At any time during the experiment, they were allowed to go back to the marked samples and re-taste them before recording their answer as to the type of water they were drinking in an unmarked cup.

C.6 Possible reasons for incorrect judgement

There may not be enough difference in taste or smell amongst the various brands of water to allow a person to detect that difference, and thus to identify correctly the origin of the water, even in the presence of identified samples of water.

C.7 Sampling plan

Three students were randomly selected from the population of students in the classroom and designated as “testers” for the experiment.

After getting their training by tasting marked samples of water, the testers were given 12 unmarked cups of water, and asked to select the brand of water (or tap water) in each cup based upon its taste and smell. Each of the four types of water was provided three times in the study, leading to 12 cups tested by all testers.

The actual identification of the unmarked cups of water, along with the responses from each of the three testers, is listed in Table C.1.

C.8 Raw data

The results from the experiment of the four trials of each type of water tested by each tester are shown in Table C.1.

Table C.1 — Results from the experiment

Cup	Actual	Tester 1	Tester 2	Tester 3	Correct percentage
1	Brand A	Brand A	Tap water	Brand B	33
2	Tap water	Brand C	Brand A	Tap water	33
3	Brand B	Brand B	Brand B	Brand A	67
4	Brand C	Brand B	Brand B	Brand A	0
5	Brand B	Tap water	Tap water	Brand C	0
6	Tap water	Brand C	Brand C	Tap water	33
7	Brand A	Brand B	Brand B	Brand C	0
8	Brand C	Tap water	Brand A	Brand B	0
9	Tap water	Tap water	Tap water	Brand C	67
10	Brand A	Brand A	Brand A	Brand A	100
11	Brand B	Brand A	Brand C	Brand C	0
12	Brand C	Brand B	Brand B	Brand C	33

C.9 Attribute agreement analysis — Within testers

C.9.1 General

AAA in SAS JMP 9[5] is adopted to assess the consistency and accuracy of subjective classifications by examining the results within testers, between testers, and against the standard.

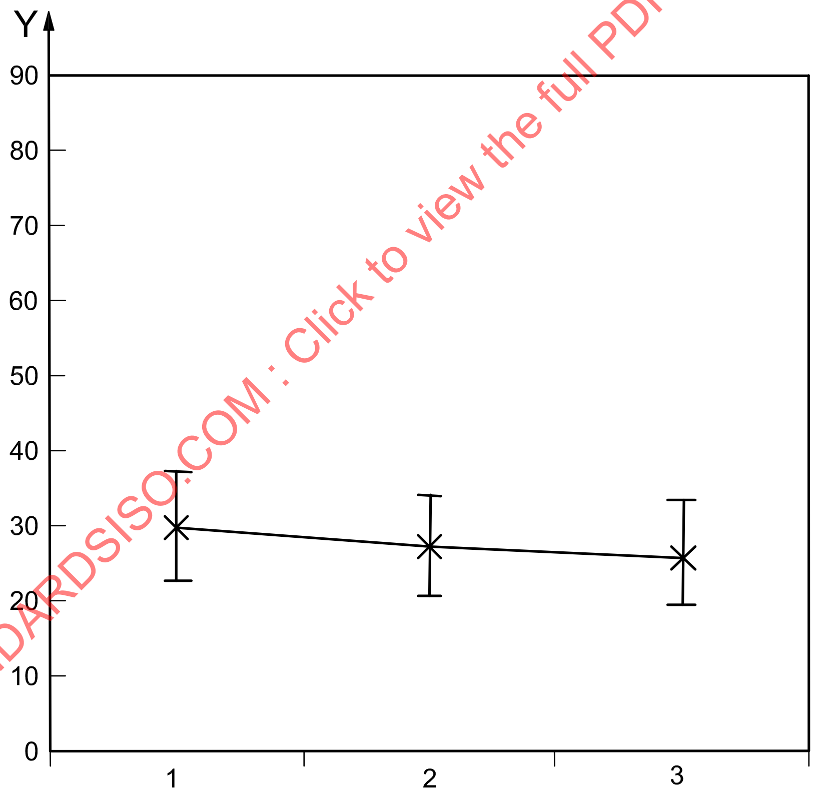
C.9.2 Consistency of each tester across trials

Table C.2 shows that none of the four types of water was recognized consistently by any tester.

Table C.2 — Agreement within testers

Rater	Number inspected	Number matched	Rater score	95 % Lower CI	95 % Upper CI
Tester 1	4	0	0,000 0	0,000 0	48,989 1
Tester 2	4	0	0,000 0	0,000 0	48,989 1
Tester 3	4	0	0,000 0	0,000 0	48,989 1

Figure C.1 and Table C.3 provide information on the percentage agreement within each tester. None of the testers seems to be consistent in their ability to identify a specific type of water.



Key

Y percentage of agreement

1 tester 1

2 tester 2

3 tester 3

Figure C.1 — Percentage agreement within each tester

[5] SAS JMP is the trade name of a product supplied by the SAS Institute, Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product named.

Table C.3 — Percentage agreement within each tester with confidence intervals

Rater	Percentage agreement	95 % Lower CI	95 % Upper CI
Tester 1	29,761 9	23,333 7	37,104 0
Tester 2	27,381 0	21,649 3	33,972 2
Tester 3	26,190 5	20,619 4	32,647 7

C.9.3 Effectiveness of each tester across trials (against standard)

The ability to effectively recognize a specific type of water is indicated in Table C.4.

Table C.4 — Effectiveness of each tester

Rater	Effectiveness	95 % Lower CI	95 % Upper CI	Error rate
Tester 1	33,333 3	13,812 0	60,937 8	0,666 7
Tester 2	25,000 0	8,894 2	53,230 5	0,750 0
Tester 3	33,333 3	13,812 0	60,937 8	0,666 7

Kappa statistics are given in Table C.5 for a comparison of each tester with the standard. None of the testers is effective in recognizing a type of water accurately.

Table C.5 — Kappa statistics of each tester against standard

Rater	Compared with standard	Kappa	Standard error
Tester 1	Actual	0,111 1	0,177 5
Tester 2	Actual	0,000 0	0,162 0
Tester 3	Actual	0,111 1	0,184 8

C.10 Attribute agreement analysis — Between testers

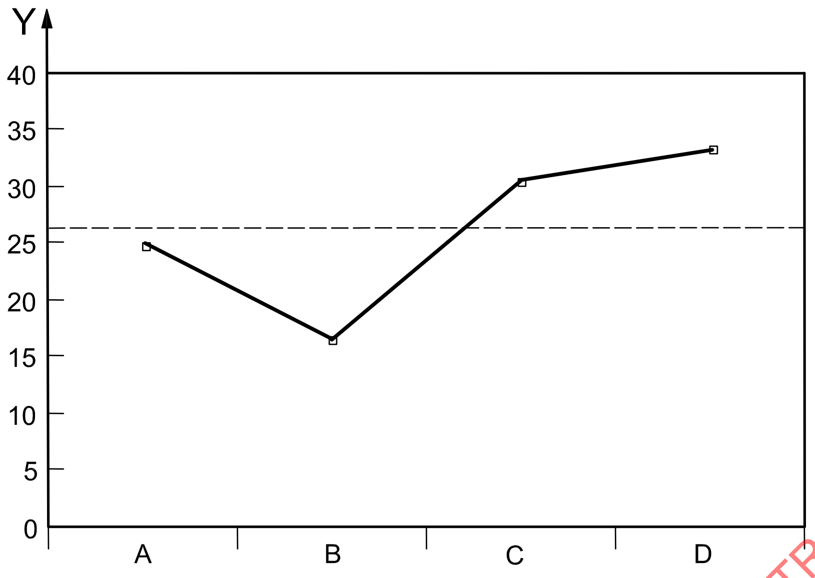
C.10.1 Consistency across testers (not against standard)

Table C.6 indicates that none of the four water types are identified consistently across testers. This is not surprising as none of the water types had been identified consistently by any single tester in Table C.2.

Table C.6 — Agreement across testers

Number inspected	Number matched	Percentage agreement	95 % Lower CI	95 % Upper CI
4	0	0,000 0	0,000 0	48,989

Figure C.2 provides the percentage agreement between testers on identifying each type of water. The highest percentage agreement is on the tap water, but it is still a low number (below 35 %). The average agreement for all types of water combined is 26 %.



Key

Y percentage of agreement

A brand A

B brand B

C brand C

D tap water

Figure C.2 — Percentage agreement between testers

Table C.7 provides kappa statistics between pairs of testers. Tester 1 and tester 2 identify eight unmarked cups out of 12 in a similar manner, leading to a kappa statistic of 0,547 2. This is better than the kappa statistics against tester 3, but still quite poor.

Table C.7 — Kappa statistics of each tester against another tester

Rater	Compared with rater	Kappa	Standard error
Tester 1	Tester 2	0,547 2	0,177 0
Tester 2	Tester 3	–0,189	0,111 5
Tester 3	Tester 3	–0,081	0,131 4

Table C.8 provides kappa statistics to assess the agreement across types of water and testers. All types of water perform poorly.

Table C.8 — Kappa statistics of each tester against each other

Category	Kappa	Standard error
Brand A	–0,037	0,083 3
Brand B	0,100 0	0,083 3
Brand C	0,000 0	0,083 3
Tap water	–0,004	0,083 3
Overall	0,016 5	0,048 2

C.10.2 Accuracy of testers against standard

Table C.9 provides counts of misclassification of types of water across trials and testers. Brand A and tap water perform better than brand B and brand C, but all types of water offer challenges to the testers and lead to a high percentage of misclassification.

Table C.9 — Misclassifications

Level	Standard			
	Brand A	Brand B	Brand C	Tap water
Brand A	—	2	2	1
Brand B	3	—	5	0
Brand C	1	3	—	4
Tap water	1	2	1	—
Other	0	0	0	0

C.11 Conclusions

AAA indicates that students are not able to consistently recognize the taste and smell of tap water or of any of the brands of bottled water. This conclusion has a big impact on answering the question that prompted the AAA in the first place, namely whether people can detect differences among specific brands of bottled water as well as between these brands and tap water. If the brands of bottled water selected for the experiment are “representative” of the population of brands, then it is unlikely that people can differentiate between bottled water and tap water.

Annex D
(informative)

Thermistor defects

D.1 General

In order to protect a thermistor (heat-sensitive resistance) from its operating atmosphere, humidity, chemical attack, and contact corrosion, the thermistor chip is often coated with protective conformal coatings that assist in assuring good mechanical integrity of the device. The coating is vitreous. During the encapsulation process, there are seven potential defects: lead outside (LO); glass crack (GC); high temperature (HT); chip turnover (CT); bubble inside glass (Bub); incomplete sealing (PS); contamination inside unit (Con). In the measurement phase, appraisers inspect the thermistor chips and subjectively classify them as passing or failing due to one of the defects mentioned above. Therefore, the experience of the appraisers and the training they have been given are of huge importance. The objective of this study is to evaluate the consistency and accuracy of the attribute measurement system.

D.2 Response variable

The response variable is nominal data, which fall into eight categories and have no natural ordering.

D.3 Standard attribute

Standard attribute (the correct rating) is given in this case.

D.4 Possible reasons for incorrect judgement

Because judgement is based on visual inspection, sometimes testers confuse similar defects, e.g. PS and Con, GC and HT.

D.5 Sampling plan

To assess the consistency and accuracy of ratings, three appraisers, randomly selected from a group with the same introductory training and similar experience, judged thermistor quality on two series of 20 samples. Thermistor samples were randomly presented.

D.6 Raw data

Table D.1 lists the raw data used in the AAA.

D.7 Attribute agreement analysis

AAA in Minitab 15 is adopted to assess the consistency and accuracy of subjective classifications by examining the results within appraisers, between appraisers, and against the standard. AAA output consists of session window and graph window results.

The session window includes the following types of agreement:

a) within appraiser: it shows whether each appraiser judges samples consistently across trials, in other words, whether the appraiser gives the same rating to the same sample each time;

b) between appraisers: it shows whether appraisers' ratings agree with each other, i.e. whether different appraisers give the same rating to the same sample.

Table D.1 — Inspection results of thermistors and standard attribute

Sample	Standard	Appraiser A		Appraiser B		Appraiser C	
		Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2
1	Pass	Pass	Pass	Pass	Pass	Pass	Pass
2	Pass	Pass	Pass	Pass	Pass	Pass	Pass
3	PS	PS	PS	PS	PS	Con	PS
4	GC	GC	GC	GC	GC	GC	GC
5	Pass	Pass	Pass	Pass	Pass	Pass	Pass
6	HT	GC	GC	HT	HT	HT	HT
7	Pass	Pass	Pass	Pass	Pass	Pass	Pass
8	GC	GC	GC	GC	HT	GC	GC
9	Pass	Pass	Pass	Pass	Pass	Pass	Pass
10	CT	CT	CT	Pass	CT	CT	CT
11	Pass	Pass	Pass	Pass	Pass	Pass	Pass
12	LO	LO	LO	LO	Pass	LO	LO
13	Pass	Pass	Pass	Pass	Pass	Pass	Pass
14	Con	Con	Con	Con	Pass	Con	PS
15	Pass	Pass	Pass	Pass	Pass	Pass	Pass
16	Pass	Pass	Pass	Pass	Pass	Pass	Pass
17	GC	GC	GC	Pass	GC	GC	GC
18	PS	PS	PS	PS	PS	PS	PS
19	Pass	Pass	Pass	Pass	Pass	PS	Pass
20	Bub	Bub	Pass	Bub	Bub	Bub	Bub

Since the standard attribute (the correct rating) is given in this case, the session window output also includes two additional types of agreement:

c) each appraiser vs. standard: it shows how well each appraiser's assessment of each sample matches with the standard, i.e. whether each rating of the same appraiser agrees with the standard rating;

d) all appraisers vs. standard: it shows how well responses of all appraisers agree with the known standard when they are combined.

For each type of agreement, the session window output includes assessment agreement and Fleiss's kappa statistics to assess the consistency and accuracy of the appraisers' responses.

D.7.1 Agreement within appraisers

The within appraisers table in the session window assists in answering whether each appraiser rated the thermistors consistently across trials.

Each appraiser inspected 20 thermistors (number inspected) twice. Table D.2 shows how well each appraiser agreed with himself/herself during two trials. Appraiser 1 matched 19 out of 20 thermistors (95 %). Appraiser 2 matched 15 out of 20 thermistors (75 %), and appraiser 3 matched 17 out of 20 thermistors (85 %).

For appraiser 1, the 95 % confidence interval (CI) for percentage matched is 75,13 % to 99,87 %. For the other two appraisers, the 95 % confidence interval for percentage matched is 50,90 % to 91,34 % and 62,11 % to 96,79 % respectively.

Table D.2 — Percentages of the assessment agreement within appraisers

Appraiser	Number inspected	Number matched[a]	Percentage	95 % CI
1	20	19	95,00	(75,13, 99,87)
2	20	15	75,00	(50,90, 91,34)
3	20	17	85,00	(62,11, 96,79)

[a] Number of times that an appraiser agrees with him/herself across trials.

To evaluate the consistency of each appraiser's ratings across trials, the kappa statistic can also be used within appraisers.

There are two main types of kappa statistic: Cohen's kappa is based on the two-way contingency table; while Fleiss's kappa is based on matched pairs. The two approaches treat the selection of appraisers differently when calculating the probability of agreement by chance. Cohen's kappa assumes that the appraiser(s) are specifically chosen and are fixed, whereas Fleiss's kappa assumes that the appraiser(s) are selected at random from a group of available appraisers. This leads to two different methods of estimating the probability. In this case, three appraisers were randomly selected from the whole group, thus it is not appropriate to employ Cohen's kappa to assess agreement. In the following, only Fleiss's kappa is considered.

Generally speaking, the higher the value of kappa, the stronger the agreement within appraisers. If $\kappa = 1$, this indicates perfect agreement (consistency). If $\kappa = -1$, this indicates perfect disagreement. If $\kappa = 0$, the agreement of the ratings is the same as that expected by chance. In general, kappa values above 0,9 are considered excellent. Kappa values less than 0,7 indicate that the rating system (or the service quality) needs improvement and those less than 0,4 indicate the measurement system capability is inadequate. Typically a kappa value of at least 0,70 is required, but kappa values close to 0,90 are preferred.

According to Table D.3, 0,924 95 is the overall Fleiss's kappa statistic for appraiser 1, which is considered good; for appraiser 2 the overall kappa is 0,59, which is not acceptable. The overall kappa for appraiser 3 is 0,79, which is acceptable.

Table D.3 also provides kappa statistics for each appraiser by defect. For example, the kappa value for appraiser 1 for defect category “Bub” is –0,02, which indicates appraiser 1 was not consistent across both trials when he/she categorized this defect category. Further observation reveals that appraiser 1 categorized the same unit as “Bub” and “Pass” in the two series of samples. This indicates appraiser 1 may be confused between categories “Bub” and “Pass”. This information would be helpful to an analyst wanting to improve the measurement system.

The p-value represents the probability of making a Type I error, which is rejecting the null hypothesis ($\kappa = 0$, or agreement within appraiser is due to chance) when the null hypothesis is true. If the p-value of a test statistic is less than the prespecified significance level (alpha), for which a commonly used value is 0,05, the null hypothesis should be rejected. Because the p-values for the three overall Fleiss's kappa are less than 0,05, the choice to reject the null hypothesis has to be made. The response agreements are significantly different from those expected by chance. The p-values for specific categories and appraisers are also shown in Table D.3.

Table D.3 — Fleiss's kappa statistics within appraisers

Appraiser	Response	Kappa	SE Kappa	Z	*p*-**Value** (vs. >0)
1	Bub	–0,025 64	0,223 607	–0,114 67	0,545 6
	Con	1,000 00	0,223 607	4,472 14	0,000 0
	CT	1,000 00	0,223 607	4,472 14	0,000 0
	GC	1,000 00	0,223 607	4,472 14	0,000 0
	HT	—[a]	—[a]	—[a]	—[a]
	LO	1,000 00	0,223 607	4,472 14	0,000 0
	Pass	0,899 75	0,223 607	4,023 80	0,000 0
	PS	1,000 00	0,223 607	4,472 14	0,000 0
	Overall	0,924 95	0,124 203	7,447 12	0,000 0
2	Bub	1,000 00	0,223 607	4,472 14	0,000 0
	Con	–0,025 64	0,223 607	–0,114 67	0,545 6
	CT	–0,025 64	0,223 607	–0,114 67	0,545 6
	GC	0,444 44	0,223 607	1,987 62	0,023 4
	HT	0,639 64	0,223 607	2,860 56	0,002 1
	LO	–0,025 64	0,223 607	–0,114 67	0,545 6
	Pass	0,583 33	0,223 607	2,608 75	0,004 5
	PS	1,000 00	0,223 607	4,472 14	0,000 0
	Overall	0,590 16	0,118 732	4,970 54	0,000 0
3	Bub	1,000 00	0,223 607	4,472 14	0,000 0
	Con	–0,052 63	0,223 607	–0,235 38	0,593 0
	CT	1,000 00	0,223 607	4,472 14	0,000 0
	GC	1,000 00	0,223 607	4,472 14	0,000 0
	HT	1,000 00	0,223 607	4,472 14	0,000 0
	LO	1,000 00	0,223 607	4,472 14	0,000 0
	Pass	0,899 75	0,223 607	4,023 80	0,000 0
	PS	0,314 29	0,223 607	1,405 53	0,079 9
	Overall	0,792 75	0,109 805	7,219 58	0,000 0

[a] No or all responses across the trials equal the value, thus kappa cannot be computed.

D.7.2 Agreement of each appraiser vs. standard

It is necessary to determine how well each appraiser's assessment of each sample matches with the standard, in other words, whether each rating of the same appraiser agrees with the standard rating.

The percentages of agreement of each appraiser vs. standard and 95 % CI can be seen in Table D.4.

Table D.4 shows how well each appraiser's assessment of each sample across both trials agrees with the standard. Appraisers 1, 2, and 3 matched 90 %, 75 %, and 85 % respectively.

Table D.4 — Percentages of the assessment agreement of each appraiser vs. standard

Appraiser	Number inspected	Number matched[a]	Percentage	95 % CI
1	20	18	90,00	(68,30, 98,77)
2	20	15	75,00	(50,90, 91,34)
3	20	17	85,00	(62,11, 96,79)

[a] Number of times that each appraiser's assessments across trials agree with the known standard.

Table D.5 shows how well each appraiser matches with the standard by defect category. This information is helpful in determining whether an appraiser has trouble with any defect type when compared with the standard. For example, appraiser 1 had good overall kappa statistics; however, he/she has trouble with the defect category “Bub”.

Table D.5 — Fleiss's kappa statistics (each appraiser vs. standard)

Appraiser	Response	Kappa	SE Kappa	Z	p-Value (vs. >0)
1	Bub	0,487 18	0,158 114	3,081 2	0,001 0
	Con	1,000 00	0,158 114	6,232 46	0,000 0
	CT	1,000 00	0,158 114	6,232 46	0,000 0
	GC	0,826 84	0,158 114	5,229 4	0,000 0
	HT	−0,025 64	0,158 114	−0,162 2	0,564 4
	LO	1,000 00	0,158 114	6,232 46	0,000 0
	Pass	0,949 87	0,158 114	6,007 5	0,000 0
	PS	1,000 00	0,158 114	6,324 6	0,000 0
	Overall	0,890 15	0,082 380	10,805 3	0,000 0
2	Bub	1,000 00	0,158 114	6,324 6	0,000 0
	Con	0,487 18	0,158 114	3,081 2	0,001 0
	CT	0,487 18	0,158 114	3,081 2	0,001 0
	GC	0,771 43	0,158 114	4,878 9	0,000 0
	HT	0,819 82	0,158 114	5,185 0	0,000 0
	LO	0,487 18	0,158 114	3,081 2	0,001 0
	Pass	0,797 98	0,158 114	5,046 9	0,000 0
	PS	1,000 00	0,158 114	6,324 6	0,000 0
	Overall	0,810 75	0,081 050	10,003 0	0,000 0
3	Bub	1,000 00	0,158 114	6,324 6	0,000 0
	Con	0,307 00	0,158 114	1,941 6	0,026 1
	CT	1,000 00	0,158 114	6,324 6	0,000 0
	GC	1,000 00	0,158 114	6,324 6	0,000 0
	HT	1,000 00	0,158 114	6,324 6	0,000 0
	LO	1,000 00	0,158 114	6,324 6	0,000 0
	Pass	0,949 87	0,158 114	6,007 5	0,000 0
	PS	0,607 94	0,158 114	3,844 9	0,000 1
	Overall	0,895 50	0,078 382	11,424 9	0,000 0

D.7.3 Agreement between appraisers

Table D.6 shows the overall agreement between the appraisers (55 %).

Table D.6 — Percentages of the assessment agreement between appraisers

Number inspected	Number matched[a]	Percentage	95 % CI
20	11	55,00	(31,53, 76,94)

[a] Number of times that all appraisers' assessments agree with each other.

Table D.7 shows agreement between appraisers by defect category. The overall kappa value is 0,74; however, the results indicate there is room for improvement among some of the defect categories such as “Con” and “HT”.

Table D.7 — Fleiss's kappa statistics between appraisers

Response	Kappa	SE Kappa	Z	p-Value (vs. >0)
Bub	0,791 304	0,057 735 0	13,705 8	0,000 0
Con	0,457 391	0,057 735 0	7,922 2	0,000 0
CT	0,791 304	0,057 735 0	13,705 8	0,000 0
GC	0,764 706	0,057 735 0	13,245 1	0,000 0
HT	0,457 391	0,057 735 0	7,922 2	0,000 0
LO	0,791 304	0,057 735 0	13,705 8	0,000 0
Pass	0,799 107	0,057 735 0	13,840 9	0,000 0
PS	0,741 193	0,057 735 0	12,837 8	0,000 0
Overall	0,742 308	0,029 869 9	24,851 3	0,000 0

D.7.4 Agreement of all appraisers vs. standard

Table D.8 lists agreement with standard by defect category when responses of all appraisers are combined. The overall values of kappa in Table D.9 indicate that the measurement system is acceptable; however, it can be further improved.

Table D.8 — Percentages of the assessment agreement of all appraisers vs. standard

Number inspected	Number matched[a]	Percentage	95 % CI
20	11	55,00	(31,53, 76,94)

[a] Number of times that all appraisers' assessments agree with the known standard.

Table D.9 — Fleiss's kappa statistics of all appraisers vs. standard

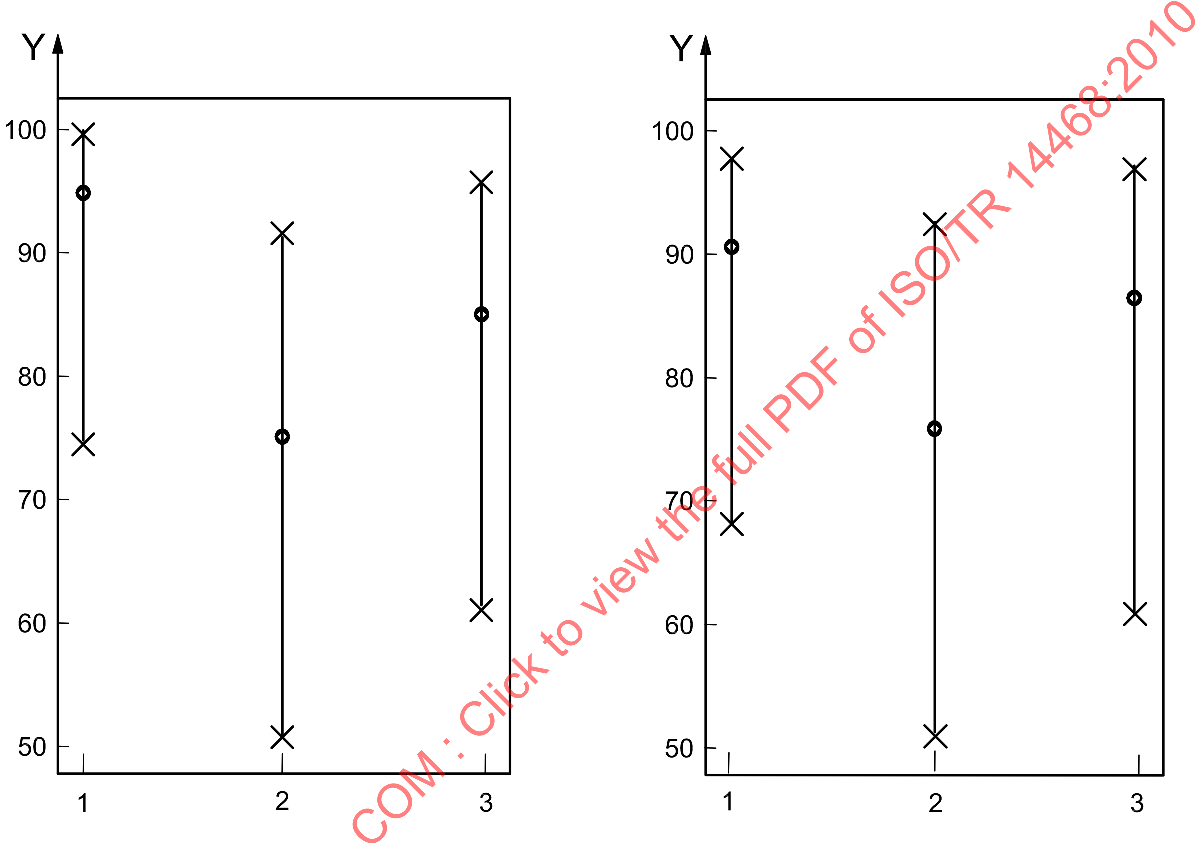
Response	Kappa	SE Kappa	Z	p-Value (vs. >0)
Bub	0,829 060	0,091 287 1	9,081 9	0,000 0
Con	0,598 060	0,091 287 1	6,551 4	0,000 0
CT	0,829 060	0,091 287 1	9,081 9	0,000 0
GC	0,866 089	0,091 287 1	9,487 5	0,000 0
HT	0,598 060	0,091 287 1	6,551 4	0,000 0
LO	0,829 060	0,091 287 1	9,081 9	0,000 0
Pass	0,899 243	0,091 287 1	9,850 7	0,000 0
PS	0,869 312	0,091 287 1	9,522 8	0,000 0
Overall	0,865 467	0,046 546 7	18,593 5	0,000 0

D.7.5 Figures of agreement assessment

The graph window also outputs two graphs: percentages of the assessment agreement and 95 % CI within appraisers on the left-hand side, percentages of the assessment agreement and 95 % CI of all appraisers vs. standard on the right (see Figure D.1).

Figure D.1 a) shows the consistency of each appraiser's ratings, while Figure D.1 b) also shows consistency and accuracy. The filled circles indicate the percentage matched and the lines joining the data points indicate a 95,0 % confidence interval.

Note that the percentage of agreement drops from 95 % to 90 % from Figure D.1 a) to b).



a) Within appraiser

b) Appraiser vs. standard

Key

Y percentage of agreement

1 appraiser 1

2 appraiser 2

3 appraiser 3

Figure D.1 — Percentage agreement of the assessment and 95 % CI

D.8 Conclusions

There are several areas where measurement system could be improved.

a) Appraiser 2 had low kappa value when compared within him/herself across trials. This indicates appraiser 2 needs additional training.

b) Each appraiser had trouble with certain defect categories when comparing within him/herself or with the standard. The definitions of defect categories should be revised and reviewed with appraisers.

c) Finally, the overall system, e.g. defect definitions, inspection procedures, and training, requires examination for further opportunities for improvement.

Annex E
(informative)

Assessment of level of disability following a stroke[6)]

E.1 General

The modified Rankin score (mRS) is a scoring measurement method that is used to assess the level of disability incurred following an acute stroke. It is traditionally based on interviews (face-to-face or telephone), which precludes its use in retrospective studies such as case control studies.

The objective of this study is to assess the consistency and accuracy of the mRS measure derived solely from case-records, thus without interviews, against the traditional mRS scores obtained from interviews.

E.2 Description of the experiment

A random sampling process was used to select 46 patients from a trial population of patients discharged from the acute stroke unit of a hospital. This trial population consisted of all outpatients sequentially discharged over a period of time from the unit and who agreed to be videotaped during the interview conducted to derive their mRS score (according to the traditional procedure). Doctors conducting the interviews were not aware that their notes would be used later on by different doctors as the sole basis to assess a case-record mRS score.

Two different assessors independently derived mRS scores from each case-record in the random sample, without knowing the actual mRS scores that had been derived initially from the interview.

E.3 Response variable

The response variable is the case-record mRS score given by each one of the two assessors. These data are of nominal type, with order (i.e. misclassifying a record which has a traditional score of 5 as a score of 1 is far more serious than misclassifying it as a score of 4).

E.4 Standard attribute

There is no “true” mRS score. The “standard mRS” score used to assess the consistency and accuracy of the case-record mRS score was based on a consensus score derived by a board of seven certified doctors who reviewed the videotapes done during the interview and agreed upon the “standard mRS score”.

E.5 Measurement method

Figure E.1 is a schematic diagram of the measurement process.

[6)] This example was supplied by Dr Terence J. Quinn, MRCP, Gardiner Institute of Cardiovascular and Medical Sciences, University of Glasgow, UK. It is adapted from Reference [6].