SWGFAST

Scientific Working Group on Friction Ridge Analysis, Study and Technology

DRAFT FOR COMMENT

STANDARD FOR THE DEFINITION AND MEASUREMENT OF RATES OF ERRORS AND NON-CONSENSUS DECISIONS IN FRICTION RIDGE EXAMINATION

(LATENT/TENPRINT)

Preamble

This document defines the concepts of error rates and non-consensus decisions as they pertain to friction ridge examination, presents strategies for measuring them, and discusses how they should be communicated.

This document defines error to mean an examiner's decision that is demonstrably wrong. The term error as it is used here is limited to individualization or exclusion conclusions that contradict ground truth or have been respectively declared exclusion or individualization by a consensus. A non-consensus decision is a determination or conclusion that cannot be assessed as an error, but conflicts with the consensus for that decision. For example, an inconclusive decision cannot be proved correct or incorrect, but can be considered non-consensus.

This standard does not mandate the measurement of rates of errors and non-consensus decisions, but prescribes terminology and approaches to be used when such rates are discussed and measured. This standard does not address the implications of errors or non-consensus decisions. If rates of errors or non-consensus decisions pertaining to friction ridge examination are presented, they must be defined in compliance with the content of this document.

1 Scope

This standard applies to the measurement of the rates of error and non-consensus decisions in all friction ridge examinations. It addresses false positives, false negatives, and non-consensus decisions resulting from the application of the ACE or ACE-V process.

2 Terminology

2.1 Definition of source attribution terms

2.1.1 Source

An area of friction ridge skin from an individual from which an impression originated.

2.1.2 Ground truth

Definitive knowledge of the actual source of an impression[1]. Ground truth cannot be determined by consensus based on friction ridge examination alone, but requires controlled collection of impressions from known individuals. This definition of ground truth cannot be satisfied with prints from operational casework. A properly administered proficiency test uses ground truth as its basis.

2.1.3 Consensus determination or conclusion

Agreement reflecting the collective judgment of a group of examiners trained to competency when making determinations or conclusions with respect to one or more impressions.

Consensus can be obtained by means determined by the testing body, such as simple majority rule of the consensus panel, qualified majority, or unanimity. The size of the consensus panel and the method used to reach consensus needs to be documented. The procedures on which consensus is based must be defined prior to the evaluation.

An audit of casework may use consensus to determine appropriate decisions or conclusions.

2.1.4 Mated

Mated impressions describe impressions intentionally collected to originate from the same source, and used for the purpose of measuring error rates. Whenever the ground truth source is not known, consensus may be used as a surrogate for the determination of the mating status of the impressions[2].

2.1.5 Non-mated

Non-mated impressions describe impressions intentionally collected to originate from different sources, and used for the purpose of measuring error rates. Whenever the ground truth source is not known, consensus may be used as a surrogate for the determination of the mating status of the impressions[3].

2.2 Definition of the types of error and error rates

An error is a conclusion reached by an examiner that contradicts the mating status of two impressions, and therefore is provably wrong (compare with non-consensus decision – see 2.3).

2.2.1 Erroneous individualization (also known as false positive or type I error)

The incorrect determination that two areas of friction ridge impressions are mated. Rates that can be computed from erroneous individualizations are:

---

[1] The term “ground truth” can also be used at the feature level. In that case, it means the definitive knowledge that all marked features correspond between two impressions.
[2] An individualization decision is in essence the determination by an examiner that two impressions are mated.
[3] An exclusion decision is in essence the determination by an examiner that two impressions are not mated.

2.2.1.1 False positive rate (FPR)

The proportion of the comparisons between non-mated prints that result in an erroneous individualization conclusion.

2.2.1.2 Positive predictive value (PPV)

The proportion of individualization decisions that are correct[4].

2.2.2 Erroneous exclusion (also known as false negative, or type II error)

The incorrect determination that two areas of friction ridge impressions are non-mated. Rates that can be computed from erroneous exclusions are:

2.2.2.1 False negative rate (FNR)

The proportion of the comparisons between mated prints that result in an erroneous exclusion conclusion.

2.2.2.2 Negative predictive value (NPV)

The proportion of exclusion determinations that are correct.

2.3 Definition of additional non-consensus decisions

A non-consensus decision is a determination or conclusion reached by an examiner at any step of the examination process that cannot be assessed as an error, but conflicts with the consensus for that decision. For example, whether or not the examination of a pair of mated impressions should result in an individualization or inconclusive decision cannot be defined in terms of error but only with respect to a consensus.

Measuring non-consensus decisions provides a means to quantify whether examiners are being overly cautious or aggressive in their decisions.

Note that non-consensus decisions indicate conflict between an examiner and a consensus, rather than a simple conflict between two individuals.

2.3.1 Non-consensus determination of suitability

When an examiner's determination of suitability does not concur with consensus. Suitability determinations include non-consensus no value, and non-consensus value decisions.

2.3.1.1 Non-consensus determinations of no value

Decisions of no value that conflict with the consensus.

2.3.1.2 Non-consensus determinations of value

Decisions of value that conflict with the consensus.

2.3.2 Non-consensus inconclusive

When an examiner reaches a decision of inconclusive that conflicts with the consensus, exclusive of false positive and negative errors.

---

[4] Most statistical definitions of positive predictive value do not have to account for inconclusive or no value decisions. The positive predictive value is sometimes presented in terms of the false discovery rate, which is defined as the percentage of individualization decisions that are incorrect. In order to account for inconclusive and no value decisions, in our case the false discovery rate is mathematically equal to 1 minus the positive predictive value minus the non-consensus individualization rate.

2.3.3 Non-consensus individualization conclusion

When an examiner reaches a decision of individualization that conflicts with the consensus, exclusive of false positive errors.

This type of non-consensus conclusion is different from the errors defined in 2.2, since the examiner is not necessarily wrong in absolute terms.

2.3.4 Non-consensus exclusion conclusion

When an examiner reaches a decision of exclusion that conflicts with the consensus, exclusive of false negative errors.

This type of non-consensus conclusion is different from the errors defined in 2.2, since the examiner is not necessarily wrong in absolute terms.

2.4 Other measures of errors and non-consensus decisions

2.4.1 Missed individualization

The failure to make an individualization when in fact both friction ridge impressions are mated (includes false negative, non-consensus inconclusive and non-consensus no value).

2.4.2 Missed exclusion

The failure to make an exclusion when in fact the friction ridge impressions are non-mated (includes false positive, non-consensus inconclusive and non-consensus no value).

3 Measurement of the rates of errors and non-consensus decisions

3.1 Rates of errors and non-consensus decisions can be measured at various resolutions based on how examiners are grouped:

3.1.1 Individual rates of errors and non-consensus decisions

The rates of errors and non-consensus decisions measured for an individual examiner pertain directly to the question of the accuracy of a specific friction ridge examination and may be the most relevant to present in court.

3.1.2 Organizational rates of errors and non-consensus decisions

The rates of errors and non-consensus decisions measured for a defined group of examiners, such as a unit, an agency, or a corporation. These rates are relevant to provide estimates of performance across an organization. They reflect the effectiveness of and adherence to standard operating procedures, proficiency testing policy, and continuing education and training programs at the organizational level.

3.1.3 Categorical rates of errors and non-consensus decisions

The rates of errors and non-consensus decisions measured for a representative group of examiners sharing one or more characteristics, such as training, experience, certification, type of cases, or caseload. These rates are relevant to provide estimates of performance across homogenous groups of examiners. Community rates of errors and non-consensus decisions

The average rates of errors and non-consensus decisions measured for the fingerprint examiner community. These rates are relevant to provide a general estimate of the accuracy of friction ridge examination, but are less pertinent when addressing the accuracy of specific examinations.

3.2 Practical considerations

3.2.1 Rates of errors and non-consensus decisions can be measured by testing the proficiency of any of the groups of fingerprint examiners described in 3.1 when examining unknown and known impressions through the ACE process. It is also possible to assess these rates after ACE-V, except when measuring rates for individual examiners.

3.2.2 The level of resources necessary to measure rates of errors and non-consensus decisions at the various resolutions defined in section 3.1 will depend on the accuracy and robustness desired for each resolution.

Because individual error rates are specific to an examiner, the number of observations required to accurately determine these rates for that examiner may be impractically large. In addition, measuring individual rates and accounting for changes in expertise over time requires extensive and repeated testing of that individual.

Organizational, categorical and community rates of errors and non-consensus decisions are averaged across multiple examiners and may be more robust to changes in personnel, and therefore require fewer observations per individual than when measuring individual rates.

Accurate measurement of community rates will require adequate representation of all categories and organizations involved in the examination of friction ridges, as well as the development of standardized tests and examination procedures, the leadership of professional organizations to administer them and encouragement to participate from the organizations described in 3.1.2.

3.2.3 Two types of tests can be designed, with respect to how the origin of the impressions is determined: ground truth, or consensus. All rates of errors and of non-consensus decisions can be measured by both tests. This is not meant to suggest that all of these need be measured, or that there is any particular deficiency if estimates of these are unknown. It is expected that in any particular study or context, only some of the rates defined in section 2 will be relevant.

3.2.4 Measuring rates when the ground truth is known

This test measures rates using a dataset of pairs of impressions of known origin collected under controlled conditions, so that the measurement of error rates relies on the ground truth, while the measurement of the rates of non-consensus decisions relies on the consensus estimation of the truth.

3.2.4.1 The measurement of the error rates will be affected by the proportions of pairs of mated and non-mated impressions. These relative proportions depend on the type of casework or operating procedures simulated. For example, AFIS casework has a very small proportion of pairs of mated impressions when considering all possible candidates.

3.2.4.2 The measurement of the error rates will be affected based on whether or not the test allows examiners to decide on the suitability of the impressions, or that comparisons are inconclusive.

3.2.4.3 The difficulty of the test will depend on the selection of the impressions. The number of non-consensus decisions of suitability, non-consensus inconclusive decisions, non-consensus individualization conclusions and erroneous exclusion conclusions will depend on the quantity and quality of information present in the mated impressions.
The number of non-consensus decisions of suitability, non-consensus inconclusive decisions, non-consensus exclusion conclusions and erroneous individualization

conclusions will depend on the quantity and quality of information present in the non-mated impressions, and on how the non-mated impressions were selected (e.g., selected randomly, through AFIS, or filtered by fingerprint characteristics, such as pattern classification).

3.2.4.4 The rates may be measured by presenting the examiner with a series of comparisons, each including one unknown print and one of the following: (1) one control impression; (2) the set of control impressions from one individual; (3) sets of control impressions from multiple individuals.

3.2.4.5 Benefits of using ground truth data when measuring rates of errors and non-consensus decisions:

- The expected conclusion for each comparison is known.
- This type of test permits the use of a greater variety of substrates, backgrounds, and development techniques than commonly observed in casework.

3.2.4.6 Challenges of measuring the rates when the ground truth is known:

- While the theoretical conclusion is known, the appropriate decisions of suitability and inconclusive must be estimated by consensus.
- Examiners may be aware that they are being tested.
- Careful collection and control of the data is required to ensure absolute confidence in the origin of every impression in the sample.
- It may be difficult to accurately replicate the relative proportions of the various combinations of substrates, backgrounds, and development techniques observed in casework. This is especially true when measuring the rates for the community as a whole.

3.2.5 Measuring rates when the ground truth is not known

This test measures rates using information obtained by auditing operational casework in which an estimate of ground truth is approximated by consensus.

3.2.5.1 Care should be taken that the selection of data is either random or sequential for the results to be considered truly representative of casework.

3.2.5.2 The measurement of the error rates will be affected based on whether or not the number of non-consensus decisions is recorded during the audit.

3.2.5.3 The data obtained during the audit may subsequently be used as test data to measure rates of errors and non-consensus decisions on other groups of examiners than the one audited, as described in 3.2.3.

3.2.5.4 Benefits of measuring rates when the ground truth is not known

- A greater amount of data may be available.
- Can accurately represent casework conditions.
- Examiners will not know that they are tested at the time when they perform the original examination.

3.2.5.5 Challenges of measuring rates when the ground truth is not known

- The ground truth can only be estimated through the use of review panels or case auditors.
- The accuracy of the measured rates will be limited by the fact that the rates are based on consensus estimates of ground truth, and therefore it is possible that the consensus panel will reach the same non-consensus decision or erroneous conclusion as the original examiner.
- The calculation of the rates may be complex if the number of known impressions varies substantially among the audited cases.

4 Calculating Rates of Errors and Non-consensus Decisions

4.1 General formulas

All observations taken into account for the numerators and denominators of the following formulas only relate to observations collected during the duration of the study.

4.1.1 False positive rate

The false positive rate is the ratio of the number of erroneous individualization conclusions (numerator) and the total number of comparisons made involving non-mated pairs (denominator).

4.1.2 False negative rate

The false negative rate is the ratio of the number of erroneous exclusion conclusions (numerator) and the total number of comparisons made involving mated pairs (denominator).

4.1.3 Positive predictive value

The positive predictive value is the ratio of the number of correct individualization conclusions (numerator) and the total number of individualization conclusions (denominator).

4.1.4 Negative predictive value

The negative predictive value is the ratio of the number of correct exclusion conclusions (numerator) and the total number of exclusion conclusions (denominator).

4.1.5 Rates of non-consensus determination of suitability

4.1.5.1 Rates of non-consensus determination of value

The rate of non-consensus determination of value is the ratio of the number of decisions of value that conflict with the consensus (numerator) and the total number of impressions determined to be of value (denominator).

4.1.5.2 Rates of non-consensus determination of no value

The rate of non-consensus determination of no value is the ratio of the number of decisions of no value that conflict with the consensus (numerator) and the total number of impressions determined to be of no value (denominator).

4.1.6 Rates of non-consensus inconclusive

The rate of non-consensus inconclusive is the ratio of the number of inconclusive comparisons that conflict with the consensus, excluding the false positive and negative errors (numerator) and the total number of inconclusive conclusions reached (denominator).

4.1.7 Rates of non-consensus individualization conclusions

The rate of non-consensus individualization is the ratio of the number of individualization conclusions that conflict with the consensus, excluding false positive errors (numerator) and the total number of individualization conclusions reached (denominator).

4.1.8 Rates of non-consensus exclusion conclusions

The rate of non-consensus exclusion is the ratio of the number of exclusion conclusions that conflict with the consensus, excluding false negative errors (numerator) and the total number of exclusion conclusions reached (denominator).

4.2 Considerations on the calculation of rates of errors and non-consensus decisions

4.2.1 The value of the various rates will be affected by whether or not the denominator of the formulas in section 4.1 include all comparisons made for any given unknown impression (e.g. the comparison of the unknown with (1) all impressions returned in an AFIS candidate list; (2) all 10 impressions from an known individual; (3) with a single impression). All these denominators are appropriate, but the choice of the testing body needs to be clearly noted whenever an error rate is cited. (see examples 2 and 3, in sections 4.3.2 and 4.3.3)

4.2.2 All errors and non-consensus decisions have to be considered for the calculation of the rates, regardless of whether they are the result of errors in judgment, clerical, or administrative errors.

4.2.3 The formulas presented in section 4.1 do not allow for combining the numbers of errors and non-consensus decisions because there is no simple formula that combines them.

4.2.4 Examiners who make excessive numbers of non-consensus decisions of unsuitability or inconclusive are likely to reduce their error rates, but will increase the numbers of missed individualizations and exclusions.

4.3 Examples for the calculation of rates of error and non-consensus decisions

The following simplified and hypothetical examples are designed to help understand the calculation of the rates of error and non-consensus decisions.

4.3.1 Hypothetical Example 1

A study is performed to measure the error rates and predictive values of a group of nine examiners. Each examiner in the group is presented with a series of 11 exercises. Each exercise includes one unknown and one known impression. Six of the exercises have mated pairs of impressions (ground truth is known to the researcher, but not to the examiner); the remaining five pairs do not have mated impressions (ground truth is known).

Each examiner is asked to examine the impressions using ACE and form a conclusion of either individualization or exclusion. Examiners are not allowed to decide on the suitability of the impressions, and cannot report inconclusive decisions.

The following data resulted from this hypothetical study:

Ground truth - Expected answer

		Mated						Non-mated				
Unknown #		1	2	3	4	5	6	7	8	9	10	11
Answer from Examiner #	1											
	2				X							
	3			X							X	
	4											
	5											
	6					X						
	7											
	8								X			
	9					X						

Table 1: Results of the test. A green color indicates a correct answer; a red color indicates a wrong answer

Based on these data, the following hypothetical error rates and predictive values can be calculated.

	Ground truth Expected answers	Mated	Non-mated	Total conclusions reached	Predictive values
Conclusions reached by Examiner 1	Individualization	6	0	6	Positive 6/6=1
	Exclusion	0	5	5	Negative 5/5=1
Total ground truth conclusions		6	5		
	Error rates	False negative 0/6=0	False positive 0/5=0		

Table 2: Error rates and predictive values for examiner 1.

		Ground truth Expected answers		Total conclusions reached	Predictive values
		Mated	Non-mated		
Conclusions reached by Examiner 2	Individualization	5	0	5	Positive 5/5=1
	Exclusion	1	5	6	Negative 5/6=0.83
Total ground truth conclusions		6	5		
	Error rates	False negative 1/6=0.16	False positive 0/5=0		

Table 3: Error rates and predictive values for examiner 2.

		Ground truth Expected answers		Total conclusions reached	Predictive values
		Mated	Non-mated		
Conclusions reached by Examiner 3	Individualization	5	1	6	Positive 5/6=0.83
	Exclusion	1	4	5	Negative 4/5=0.8
Total ground truth conclusions		6	5		
	Error rates	False negative 1/6=0.16	False positive 1/5=0.2		

Table 4: Error rates and predictive values for examiner 3.

		Ground truth		Total conclusions	Predictive

Expected answers				reached	values
		Mated	Non-mated		
Conclusions reached by the group	Individualization	50	2	52	Positive 50/52=0.96
	Exclusion	4	43	47	Negative 43/47=0.91
Total ground truth conclusions		54	45		
Error rates		False negative 4/54=0.074	False positive 2/45=0.044		

Table 5: Error rates and predictive values for the group.

4.3.2 Hypothetical Example 2

An audit of operational casework was performed in laboratory Oz to determine the organization rates of errors and non-consensus decisions.

A panel of 4 examiners trained to competency performed the audit. In this example, consensus was defined as at least 3 out of the 4 examiners performing the audit concurring on a decision. The auditors reviewed casework and made decisions using laboratory Oz's standard operating procedures, which were based on SWGFAST documents.

The audit was considering the operation of laboratory Oz over the past 10 years, during which time 10,000 latent impressions were examined. For the audit, the auditors selected 1,000 latent impressions. The selection of these impressions could have occurred in various ways: random selection over the entire pool, or stratified selection based on various categories, such as selecting pre-defined numbers of individualization/exclusion decisions, or equal numbers per examiner. For the purpose of this example, it is assumed that the samples were selected randomly.

In this example, comparison decisions (individualization, inconclusive, exclusion) are calculated by subject, not calculated separately per finger.[5]

The original decisions made by laboratory Oz on the selected 1,000 latent impressions were as follows:

[5] Note that calculations by subject do not specifically address errors in which an individualization is made to the correct subject but the wrong area of friction ridge skin.

Original decision	Number
No value	80
Of value but no suitable candidate	220
Exclusion	160
Inconclusive	240
Individualization	300

After auditing the impressions, the audit panel found the following results. Errors are highlighted in yellow; non-consensus decisions are highlighted in blue.

Original decision		Audit				
		No value	Of value but no candidate	Exclusion	Inconclusive	Individualization
No value	80	69	1	1	8	1
Of value but no candidate[6]	220	2	218	0	0	0
Exclusion	160	4	0	140	10	6
Inconclusive	240	23	0	7	200	10
Individualization	300	14	0	1	15	270
	1000	112	219	149	233	287

[6] "Of value but no candidate" means no subjects were compared, or all AFIS candidate images that were compared were excluded.

Description	Numerator	Denominator	Rates
False positive	1	149	0.007
Positive predictive value	270	300	0.900
False negative	6	287	0.021
Negative predictive value	140	160	0.875
Non-consensus no value	11 (1+1+8+1)	80	0.138
Non-consensus value	43 (2+4+23+14)	920 (220+160+240+300)	0.047
Non-consensus inconclusive	40 (23+7+10)	240	0.167
Non-consensus exclusion	14 (10+4)	160	0.088
Non-consensus individualization	29 (15+14)	300	0.097
Missed individualization	17 (1+6+10)	287	0.059
Missed exclusion	9 (1+7+1)	149	0.060

4.3.3 Hypothetical Example 3

An audit of operational casework was performed in laboratory Narnia to determine the organization rates of errors and non-consensus decisions. All of Narnia's casework is performed on AFIS. This example shows the effect of calculations on a per-image basis.

As in example 2, a panel of 4 examiners trained to competency performed the audit. In this example, consensus was defined as at least 3 out of the 4 examiners performing the audit concurring on a decision. The auditors reviewed casework and made decisions using laboratory Narnia's standard operating procedures, which were based on SWGFAST documents.

The audit was considering the operation of laboratory Narnia over the past year, during which time 1,000 latent impressions were searched. The original candidate list for each search was retained and used in the audit. For the audit, all 1,000 latent impressions were reviewed.

In this example, comparison/evaluation decisions (individualization, inconclusive, exclusion) are calculated by separately for each image compared.

The original decisions made by laboratory Narnia on the selected 1,000 latent impressions were as follows. In every case where the latent was of value, 20 AFIS candidate images were compared, so each latent that was of value and searched resulted in 20 comparison/evaluation decisions.

Original decision	Number
No value for AFIS search	150
Exclusion	16,000
Inconclusive	750
Individualization	250

After auditing the impressions, the audit panel found the following results. Errors are highlighted in yellow; non-consensus decisions are highlighted in blue.

Original decision		Audit consensus				
		Latent is no value	Latent is of value (not searched)	Exclusion	Inconclusive	Individualization
No value for AFIS search	150	140	10	n/a	n/a	n/a
Exclusion	16,000	n/a	n/a	15,844	150	6*
Inconclusive	750	n/a	n/a	16	730	4
Individualization	250	n/a	n/a	2*	3	245
				15,862	883	255

* One latent was erroneously individualized to an image in its candidate list when the correct subject was in the candidate list; therefore that latent resulted in one false positive and one false negative.

Description	Numerator	Denominator	Rates
False positive	2	15,862	0.0001
Positive predictive value	245	250	0.980
False negative	6	255	0.023
Negative predictive value	15,844	16,000	0.9903
Non-consensus no value	10	150	0.067
Non-consensus value	n/a	n/a	n/a
Non-consensus inconclusive	20 (16+4)	750	0.027
Non-consensus exclusion	150	16,000	0.0093
Non-consensus individualization	3	250	0.012
Missed individualization	10 (6+4)	255	0.039
Missed exclusion	18 (16+2)	15,862	0.0011

5 Interpretation and communication of the rates of errors and non-consensus decisions

5.1 The hypothetical error rates and predictive values reported in table 5 must be interpreted and communicated as follows. Rates of non-consensus decisions can be interpreted and communicated in a similar fashion.

5.1.1 False positive rate

Table 5 indicates a false positive rate of 0.044. This indicates that examiners in the tested group erroneously identified the source of an impression on average in 4.4% of the examinations that they performed under the test conditions.

5.1.2 False negative rate

Table 5 indicates a false negative rate of 0.074. This indicates that examiners in the tested group erroneously excluded the source of an impression on average in 7.4% of the examinations that they performed under the test conditions.

5.1.3 Positive predictive value

Table 5 indicates a positive predictive value of 0.96. This indicates that examiners in the tested group and under the test conditions, who report an individualization conclusion, were correct, on average, 96% of the time.

5.1.4 Negative predictive value

Table 5 indicates a negative predictive value of 0.91. This indicates that examiners in the tested group and under the test conditions, who report an exclusion conclusion, were correct, on average, 91% of the time.

5.2 Understanding rates of errors and non-consensus decisions

5.2.1 Error rates cannot be communicated using a single number. Each of the values presented in section 2 has different implications and interpretations. Rates of errors and non-consensus decisions should be reported together in order to properly interpret the effectiveness and accuracy of examiners.

For example, an examiner who has a very low rate of false exclusions, but a high rate of non-consensus no value decisions can be considered as excessively cautious; and therefore, less effective than another examiner with the same rate of false exclusions but a lower rate of non-consensus no value decisions.

5.2.2 The selection of which of the rates defined in this document will be measured and reported will depend on the information solicited.

5.2.3 The communication of these rates needs to be supported by the disclosure of the experimental conditions in which they have been measured, such as the group of examiners to which they apply (see section 3.1), the type of casework considered (see section 3.2), whether they are based on ground truth or consensus (see sections 3.2), and if consensus was used, how consensus was established (see section 2.1.3).

5.2.4 If it is determined that some errors or non-consensus decisions were administrative or clerical in nature, this should be noted and disclosed to better explain the corresponding rates.

5.2.5 These rates can be meaningfully compared between different individuals, groups of examiners or organizations, only when they were measured under similar conditions.