DRAFT FOR COMMENT

STANDARD FOR THE DEFINITION AND MEASUREMENT OF RATES OF ERRORS AND INAPPOPRIATE DECISIONS IN FRICTION RIDGE EXAMINATION

Preamble

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

This standard does not mandate the measurement of rates of errors and inappropriate decisions, but prescribes terminology and approaches to be used when such rates were measured. If rates of errors or inappropriate decisions pertaining to friction ridge examination are presented, they must comply with the content of this document.

1   Scope

This standard applies to the measurement of the rates of error and inappropriate decisions in all friction ridge examinations. It addresses false positives, false negatives, and inappropriate 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 origin of an impression at the source level[1]. Ground truth cannot be determined by consensus based on friction ridge examination alone.

---

[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.

DRAFT FOR COMMENT

2.1.3 Consensus determination or conclusion

Agreement reflecting the collective judgment of examiners 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 method used to reach consensus needs to be documented.

2.1.4 Mated

Two impressions from the same source, as determined by ground truth knowledge or consensus determination.

2.1.5 Non-mated

Two impressions not from the same source, as determined by ground truth knowledge or consensus determination.

2.2 Definition of the types of error and error rates

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:

2.2.1.1 False positive rate

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

2.2.1.2 Positive predictive value

The percentage of individualization decisions that are correct. This is sometimes presented as the false discovery rate (percentage of individualization decisions that are incorrect; mathematically expressed as 1 minus positive predictive value).

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

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

2.2.2.2 Negative predictive value

The percentage of exclusion determinations that are correct.

2.3 Definition of additional inappropriate decisions

Inappropriate decisions are determinations or conclusions that conflict with the consensus. Inappropriate decisions cannot be judged against ground truth, as defined in 2.1.2.

2.3.1 Inappropriate determination of suitability

When an examiner's determination of suitability does not concur with consensus. The rate of inappropriate determinations of suitability is the percentage of decisions of suitability that conflict with the consensus.

Standards for the Definition and Measurement of Rates of Errors and Inappropriate Decisions in Friction Ridge Examination

DRAFT FOR COMMENT

02/11/11 ver. 1.0
Posted: 3/15/11

Page 2 of 11

2.3.2 Inappropriate inconclusive

When an examiner reaches a decision of inconclusive, and it is otherwise determined that a conclusion (individualization or exclusion) should have been reached. The rate of inappropriate inconclusive decisions is the percentage of decisions of inconclusive that conflict with the consensus.

2.3.3 Inappropriate individualization conclusion

When an examiner reaches a decision of individualization, and it is otherwise determined that a decision of inconclusive should have been reached. This type of inappropriate conclusion is different from the errors defined in 2.2, since the examiner is not necessarily wrong in absolute terms.

The rate of inappropriate individualization conclusions is the percentage of conclusions of individualization that are deemed inconclusive by the consensus.

2.3.4 Inappropriate exclusion conclusion

When an examiner reaches a decision of exclusion, and it is otherwise determined that a decision of inconclusive should have been reached. This type of inappropriate conclusion is different from the errors defined in 2.2, since the examiner is not necessarily wrong in absolute terms.

The rate of inappropriate exclusion conclusions is the percentage of conclusions of exclusion that are deemed inconclusive by the consensus.

2.4 Other measures of errors and inappropriate decisions

2.4.1 Missed individualization

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

3 Measurement of the rates of errors and inappropriate decisions

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

3.1.1 Individual rates of errors and inappropriate decisions

The rates of errors and inappropriate decisions measured for an individual examiner are the most relevant, as they pertain directly to the question of the accuracy of a specific friction ridge examination. Measurement of individual rates requires extensive and repeated testing over time. Measuring individual rates may be impractical and resource-intensive.

3.1.2 Organizational rates of errors and inappropriate decisions

The rates of errors and inappropriate 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 impact of existence and quality of standard operating procedures, proficiency testing policy, and continuing education and training programs at the organizational level.

DRAFT FOR COMMENT

3.1.3 Categorical rates of errors and inappropriate decisions

The rates of errors and inappropriate 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.

3.1.4 Community rates of errors and inappropriate decisions

The average rates of errors and inappropriate 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. 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 Practical considerations

3.2.1 Rates of errors and inappropriate 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.

3.2.2 Two types of tests can be designed, with respect to how the origin of the impressions is determined. All rates of errors and of inappropriate 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.

3.2.3 Measuring rates when the ground truth is known

The first type of test can measure these rates using a dataset of pairs of impressions of known origin collected under controlled conditions.

3.2.3.1 In this type of test, the measurement of error rates relies on the ground truth, while the measurement of the rates of inappropriate decisions relies on the consensus estimation of the truth.

3.2.3.2 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.

3.2.3.3 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.

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

The number of inappropriate decisions of suitability, inappropriate inconclusive decisions, inappropriate 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).

3.2.3.5 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.

3.2.3.6 Benefits of using ground truth data when measuring rates of errors and inappropriate 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.

3.2.3.7 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.

3.2.4 Measuring rates when the ground truth is not known

The second type of test will measure these rates using information obtained by auditing operational casework in which an estimate of ground truth is approximated by consensus.

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

3.2.4.2 The measurement of the error rates will be affected based on whether or not the number of inappropriate decisions is recorded during the audit.

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

3.2.4.4 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.

3.2.4.5 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.

Standards for the Definition and Measurement of Rates of Errors and Inappropriate Decisions in Friction Ridge Examination

DRAFT FOR COMMENT

02/11/11 ver. 1.0
Posted: 3/15/11

Page 5 of 11

- The accuracy of the measured rates will be limited by the possibility that the ground truth may be incorrectly estimated if the consensus panel reaches the same inappropriate decision or erroneous conclusion as the original examiner.

- The calculation of the rates may be rendered more complex by the lack of homogeneity in the number of known impressions that are included in the audited cases.

4   Calculating Rates of Errors and Inappropriate Decisions

4.1 General formulas

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). The positive predictive value may be stated as the false discovery rate, which is 1 minus the positive predictive value.

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 inappropriate determination of suitability

The rate of inappropriate determination of suitability is the ratio of the number of decisions of suitability that conflict with the consensus (numerator) and the total number of unknown impressions considered (denominator).

4.1.6 Rates of inappropriate inconclusive

The rate of inappropriate inconclusive is the ratio of the number of inconclusive comparisons that conflict with the consensus (numerator) and the total number of conclusions reached (denominator).

4.1.7 Rates of inappropriate individualization conclusion

The rate of inappropriate individualization is the ratio of the number of individualization conclusions that conflict with the consensus (numerator) and the total number of individualization conclusions reached (denominator).

4.1.8 Rates of inappropriate exclusion conclusion

The rate of inappropriate exclusion is the ratio of the number of exclusion conclusions that conflict with the consensus (numerator) and the total number of exclusion conclusions reached (denominator).

Standards for the Definition and Measurement of Rates of Errors
and Inappropriate Decisions in Friction Ridge Examination

DRAFT FOR COMMENT

02/11/11 ver. 1.0
Posted: 3/15/11

Page 6 of 11

4.2 Considerations on the calculation of rates of errors and inappropriate decisions

4.2.1 All errors and inappropriate decisions have to be considered for the calculation of the rates, regardless of whether they are the result of clerical or administrative errors.

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

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

4.2.4 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.

4.3 Example for the calculation of error rates and predictive values

The following simplified and hypothetical example is designed to help understand the calculation of the false positive and negative error rates and the predictive values. The rates of inappropriate decisions can be measured in a similar fashion.

Hypothetical Example

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 impressions (ground truth is known); the remaining five 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:

Standards for the Definition and Measurement of Rates of Errors
and Inappropriate Decisions in Friction Ridge Examination

DRAFT FOR COMMENT

02/11/11 ver. 1.0
Posted: 3/15/11

Page 7 of 11

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		Total conclusions reached	Predictive values
		Mated	Non-mated		
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.

DRAFT FOR COMMENT

		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.

DRAFT FOR COMMENT

	Ground truth Expected answers			Total conclusions reached	Predictive 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.

5 Interpretation and communication of the rates of errors and inappropriate decisions

    5.1 The hypothetical error rates and predictive values reported in table 5 must be interpreted and communicated as follows. Rates of inappropriate 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 inappropriate decisions

        5.2.1 Error rates cannot be communicated using a single number. Each of the values presented in section 2 has a different implication.

        5.2.2 One or more of these values may be measured and reported depending on the information solicited.

Standards for the Definition and Measurement of Rates of Errors and Inappropriate Decisions in Friction Ridge Examination

DRAFT FOR COMMENT

02/11/11 ver. 1.0
Posted: 3/15/11

Page 10 of 11

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 inappropriate 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.

DRAFT FOR COMMENT