INTERNATIONAL STANDARD

ISO/IEC 19795-1

Second edition
2021-05

Information technology — Biometric performance testing and reporting —

Part 1: Principles and framework

Technologies de l'information — Essais et rapports de performance biométriques —

Partie 1: Principes et canevas

Reference number
ISO/IEC 19795-1:2021(E)

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

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

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

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see patents.iec.ch).

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

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

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 37, *Biometrics*.

This second edition cancels and replaces the first edition (ISO/IEC 19795-1:2006), which has been technically revised.

The main changes compared to the previous edition are as follows:

— Terminology is updated to follow the biometrics vocabulary of ISO/IEC 2382-37:2017;

— Additional detail is provided on testing and reporting of transaction times and computational workload, and on graphical representation of results.

A list of all parts in the ISO 19795 series can be found on the ISO website.

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

Introduction

This document is concerned solely with the scientific technical performance testing of biometric systems and devices. Technical performance testing seeks to determine error and throughput rates, with the goal of understanding and predicting the real-world error and throughput performance of biometric systems. The error rates include both false-positive and false-negative rates, as well as failure-to-enrol and failure-to-acquire rates across the test population. Throughput rates refer to the number of individuals processed per unit of time based both on computational speed and human–machine interaction. These measures are generally applicable to all biometric systems and devices. Technical performance tests that are modality-specific, for example, fingerprint scanner image quality, are not considered in this document.

The purpose of this document is to present the requirements and best scientific practices for conducting and reporting technical performance testing. It is acknowledged that technical performance testing is only one form of biometric testing. Other types of testing not considered in this document include:

— reliability, availability and maintainability;

— security, including vulnerability;

— conformance;

— safety;

— human factors, including user acceptance;

— cost/benefit;

— privacy regulation conformance.

Biometric technical performance testing can be of three types: technology, scenario and operational evaluation. Each type of test requires a different protocol and produces different types of results. Other parts of the ISO/IEC 19795 series provide specific advice and requirements for the development and use of such different test protocols. This document addresses specific philosophies and principles that can be applied over a broad range of test conditions.

Information technology — Biometric performance testing and reporting —

Part 1: Principles and framework

1 Scope

This document:

— establishes general principles for testing the performance of biometrics systems in terms of error rates and throughput rates for purposes including measurement of performance, prediction of performance, comparison of performance, and verifying conformance with specified performance requirements;

— specifies performance metrics for biometric systems;

— specifies requirements on the recording of test data and reporting of test results; and

— specifies requirements on test protocols in order to:

  — reduce bias due to inappropriate data collection or analytic procedures;

  — help achieve the best estimate of field performance for the expended effort;

  — improve understanding of the limits of applicability of the test results.

This document is applicable to empirical performance testing of biometric systems and algorithms through analysis of the comparison scores and decisions output by the system, without requiring detailed knowledge of the system's algorithms or of the underlying distribution of biometric characteristics in the population of interest.

Not within the scope of this document is the measurement of error and throughput rates for people deliberately trying to subvert the intended operation of the biometric system (e.g. by presentation attacks).

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 2382-37, *Information technology — Vocabulary — Part 37: Biometrics*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 2382-37 and the following apply.

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

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

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

3.1
test subject
individual whose biometric data is intended to be enrolled or compared as part of the evaluation

3.2
test crew
set of *test subjects* (3.1) utilized in an evaluation

3.3
target population
set of biometric data subjects of the application for which performance is being evaluated

3.4
test organization
functional entity under whose auspices the test is conducted

3.5
experimenter
individual responsible for defining, designing and analysing the test

3.6
test administrator
individual performing the testing

EXAMPLE Staff conducting enrolments or overseeing verification or *identification transactions* (3.10).

3.7
test observer
individual recording test data or monitoring the *test crew* (3.2)

3.8
enrolment attempt
sequence of one or more capture attempts with the aim of producing a biometric reference for a capture subject

Note 1 to entry: An enrolment attempt can require a specific number of capture attempts (e.g. three separate placements of a finger on a sensor within a set period), from which the highest quality sample(s) is/are selected for further processing.

3.9
enrolment transaction
one or more *enrolment attempts* (3.8) with the aim of producing a biometric reference for a capture subject

Note 1 to entry: If an enrolment attempt fails, further enrolment attempts can be performed within the same enrolment transaction until an attempt succeeds or enrolment is given up.

3.10
identification transaction
sequence of one or more capture attempts and biometric searches to find and return the biometric reference identifier(s) attributable to a single individual

3.11
channel effect
variation of the biometric sample due to sampling, noise and frequency response characteristics of the sensor and transmission channel

3.12
presentation effect
variation of the biometric sample due to the way that biometric characteristics are presented to the sensor

EXAMPLE In facial recognition, this can include pose angle; in fingerprinting, finger rotation and skin moisture. In many cases, the distinction between changes in the fundamental biometric characteristic and the presentation effects are unclear (e.g. facial expression in facial recognition or pitch change in speaker verification systems).

3.13
technology evaluation
offline (3.17) evaluation of one or more algorithms for the same biometric modality using a pre-existing or especially-collected corpus of samples

3.14
scenario evaluation
evaluation that measures end-to-end system performance in a prototype or simulated application with a *test crew* (3.2)

3.15
operational evaluation
evaluation that measures the performance of a biometric system in a specific application environment using a specific *target population* (3.3)

3.16
online
pertaining to execution of biometric enrolment or comparison directly following the biometric acquisition process

3.17
offline
pertaining to execution of biometric enrolment or comparison of stored biometric data subsequent to and disconnected from the biometric acquisition process

Note 1 to entry: Collecting a corpus of images or signals for offline enrolment and calculation of comparison scores allows greater control over which probe and reference images are to be used in any transaction.

3.18
closed-set test
test in which the *test crew* (3.2) comprises only individuals known to have a reference in the enrolment database

Note 1 to entry: Closed-set tests are a specific type of test for showing performance of identification systems in terms of a *cumulative match characteristic plot* (3.29).

3.19
failure to acquire
failure of the biometric capture and feature extraction processes to produce biometric features suitable for biometric comparison

3.20
false reject rate
FRR
proportion of verification transactions with true biometric claims erroneously rejected

3.21
false accept rate
FAR
proportion of verification transactions with false biometric claims erroneously accepted

3.22
false-negative identification rate
FNIR
FNIR(*N*, *R*, *T*)
proportion of a specified set of *identification transactions* (3.10) by capture subjects enrolled in the system for which the subject's correct reference identifier is not among those returned

Note 1 to entry: The false-negative identification rate can be expressed as a function of *N*, the number of enrolees, and of parameters of the identification process where only candidates up to *rank* (3.24) *R*, and with a candidate score greater than threshold *T* are returned to the candidate list.

3.23
false-positive identification rate
FPIR
FPIR(*N*, *T*)
proportion of *identification transactions* (3.10) by capture subjects not enrolled in the system for which a reference identifier is returned

Note 1 to entry: The false-positive identification rate can be expressed as a function of *N*, the number of enrolees, and parameters of the identification process where only candidates with a candidate score greater than threshold *T* are returned to the candidate list.

Note 2 to entry: For systems that always return a fixed number of candidates without applying a threshold on scores, FPIR is not a meaningful metric.

3.24
rank
position of a candidate in a candidate list ordered by descending similarity score

3.25
true-positive identification rate
TPIR
TPIR(*N*, *R*, *T*)
proportion of *identification transactions* (3.10) by capture subjects enrolled in the system for which the subject's correct identifier is among those returned

Note 1 to entry: The true-positive identification rate can be expressed as a function of *N*, the number of enrolees, and of parameters of the identification process where only candidates up to *rank* (3.24) *R*, and with a candidate score greater than threshold *T* are returned to the candidate list.

Note 2 to entry: TPIR(*N*, *R*, *T*) = 1 – FPIR(*N*, *R*, *T*).

3.26
selectivity
SEL(*N*, *R*, *T*)
average number of candidates returned above threshold *T* in a non-mated *identification transaction* (3.10)

Note 1 to entry: Selectivity can be expressed as a function of *N*, the number of enrolees, and of parameters of the identification process where only candidates up to *rank* (3.24) *R* and with candidate score greater than threshold *T* are returned on the candidate list.

Note 2 to entry: When *R* = *N*, SEL(*N*, *R*, *T*) is measured against the entire database.

3.27
computational workload
total computational effort of a single transaction (or set of transactions) in a biometric system, including number of intrinsic operations, execution time and memory requirements

Note 1 to entry: Computational workload is dependent on the hardware on which the biometric system is operating.

3.28
detection error trade-off
DET
relationship between false-negative and false-positive errors of a binary classification system as the discrimination threshold varies

Note 1 to entry: The DET can be represented as a DET table or as a DET plot.

Note 2 to entry: The receiver operating characteristic (ROC) curve was used in the previous edition of this document. The ROC is unified with the DET.

3.29
cumulative match characteristic plot
CMC plot
graphical presentation of results of mated searches in a closed-set identification test, plotting the *true-positive identification rate* (3.25), TPIR(*N, R,* 0), as a function of *R*

3.30
pre-selection algorithm
algorithm to reduce the number of comparisons that need to be made in an identification search of the enrolment database

3.31
pre-selection error
<pre-selection algorithm> error that occurs when the corresponding subject identifier is not in the pre-selected subset of candidates

Note 1 to entry: In binning pre-selection, pre-selection errors occur when the data subject's enrolment reference and a subsequent sample from the same biometric characteristic are placed in different partitions.

3.32
penetration rate
<pre-selection algorithm> average proportion of the total number of references that are pre-selected

4 Abbreviated terms

API	application programming interface
CMC	cumulative match characteristic
FAR	false accept rate
FTAR	failure-to-acquire rate
FTCR	failure-to-capture rate
FTER	failure-to-enrol rate
FTXR	failure-to-extract rate
FNIR	false-negative identification rate
FPIR	false-positive identification rate
FRR	false reject rate
GFAR	generalized false accept rate
GFRR	generalized false reject rate

PIN	personal identification number
RFID	radio frequency identification
ROC	receiving operating characteristic
SDK	software developer's kit
SEL	selectivity
TPIR	true-positive identification rate

5 Conformance

To conform to this document, a biometric performance test shall be planned, executed and reported in accordance the requirements contained in Clauses 7 through 12.

6 General biometric system

6.1 Conceptual representation of general biometric system

Given the variety of applications and technologies, it can seem difficult to draw any generalizations about biometric systems. All such systems, however, have many elements in common. Captured biometric samples are acquired from a subject by a biometric capture device and are sent to a processor that extracts the distinctive but repeatable measures of each sample (the biometric features), discarding all other components. The resulting features may be stored in the biometric enrolment database as a biometric reference. In other cases, the sample itself (without feature extraction) may be stored as the reference. A subsequent query or probe biometric sample can be compared to a specific reference, to many references, or to all references already in the database to determine if there is a match. A decision regarding the biometric claim is made based upon the similarities or dissimilarities between the features of the biometric probe and those of the reference or references compared.

Figure 1 illustrates the information flow within a general biometric system consisting of data capture, signal processing, data storage, comparison and decision subsystems. This diagram illustrates both enrolment and the operation of verification and identification systems.

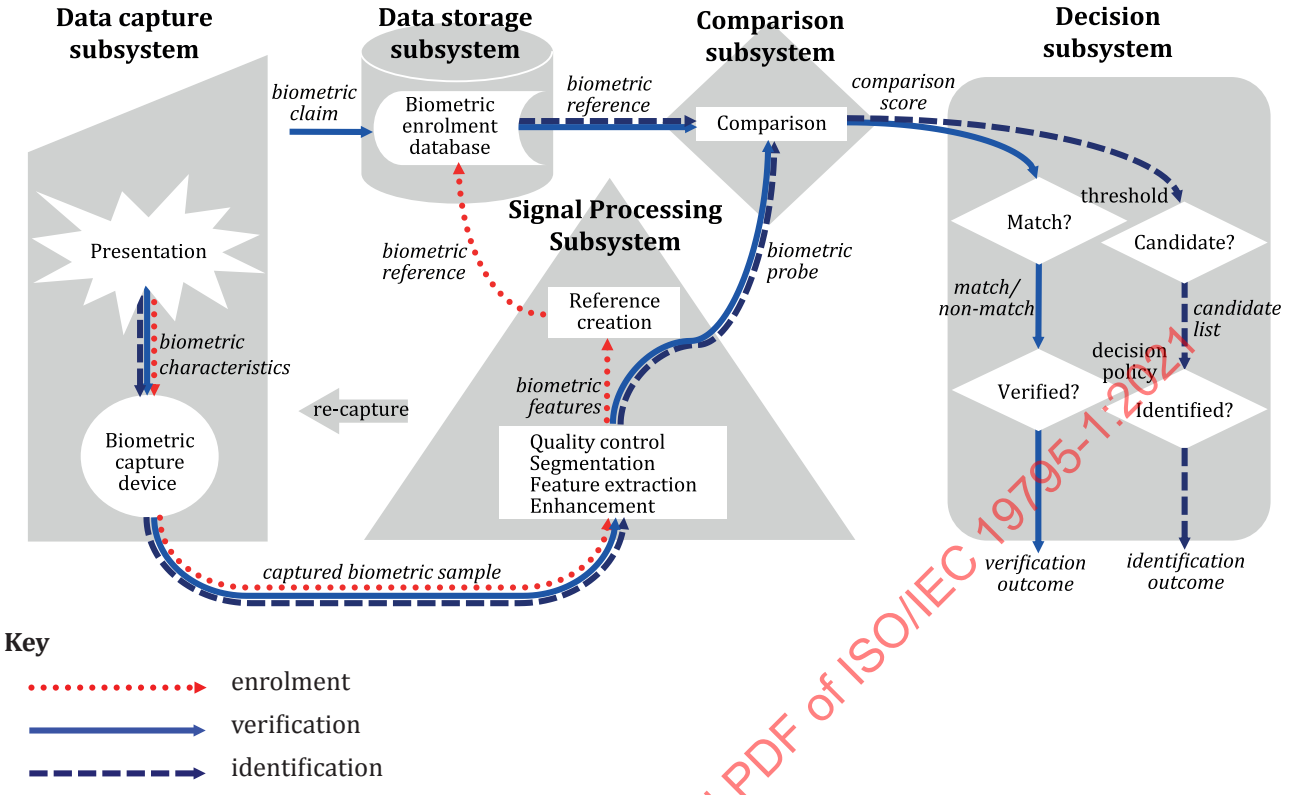


Figure 1 — Components of a general biometric system

The following subclauses describe each of these subsystems in more detail. However, it should be noted that in any implemented system, some of these conceptual components may be absent, or may not have a direct correspondence with a physical or software entity.

6.2 Conceptual components of a general biometric system

6.2.1 Data capture subsystem

The data capture subsystem collects an image or signal of a subject's biometric characteristics presented to the biometric capture device, and outputs this image or signal as a captured biometric sample.

6.2.2 Transmission subsystem

The transmission subsystem (not always present or visibly present in a biometric system) transmits samples, features, probes, references, comparison scores and outcomes between different subsystems. The captured biometric sample may be compressed and/or encrypted before transmission and expanded and/or decrypted before use. A captured biometric sample may be altered in transmission due to noise in the transmission channel as well as losses in the compression/expansion process. Data may be transmitted using standard biometric data interchange formats, and cryptographic techniques may be used to protect the authenticity, integrity, and confidentiality of stored and transmitted biometric data.

NOTE The transmission subsystem is not portrayed in Figure 1.

6.2.3 Signal processing subsystem

Signal processing includes processes such as:

— enhancement, i.e. improving the quality and clarity of the captured biometric sample;

— segmentation, i.e. locating the signal of the subject's biometric characteristics within the captured biometric sample;

— feature extraction, i.e. deriving the subject's repeatable and distinctive measures from the captured biometric sample; and

— quality control, i.e. assessing the suitability of samples, features, references, etc. and possibly affecting other processes, such as returning control to the data capture subsystem to collect further samples (recapture), or modifying parameters for segmentation, feature extraction, or comparison.

In the case of enrolment, the signal processing subsystem creates a biometric reference. Sometimes the enrolment process requires features from several presentations of the individual's biometric characteristics. Sometimes the reference comprises just the features, in which case the reference may be called a "template". Sometimes the reference comprises just the sample, in which case feature extraction from the reference occurs immediately before comparison.

In the case of verification and identification, the signal processing subsystem creates a biometric probe.

Sequencing and iteration of the above-mentioned processes are determined by the specifics of each system.

6.2.4 Data storage subsystem

References are stored within an enrolment database held in the data storage subsystem. Each reference may be associated with some details of the enrolled subject or the enrolment process. It should be noted that prior to being stored in the enrolment database, references may be reformatted into a biometric data interchange format. References may be stored within a biometric capture device, on a portable medium such as a smart card, locally such as on a personal computer or local server, in a central database, or in the 'cloud'.

6.2.5 Comparison subsystem

In the comparison subsystem, probes are compared against one or more references and comparison scores are passed to the decision subsystem. The comparison scores indicate the similarities or dissimilarities between the probe(s) and reference(s) compared. For verification, a single specific biometric claim would lead to a single comparison score. For identification, many or all references may be compared with the probes and output a comparison score for each comparison.

6.2.6 Decision subsystem

The decision subsystem uses the comparison scores generated from one or more biometric comparisons to provide the decision outcome for a verification or identification transaction.

In the case of verification, the probes are considered to match a compared reference when (assuming that higher scores correspond to greater similarity) the comparison score exceeds a specified threshold. A biometric claim can then be verified on the basis of the decision policy, which may allow or require multiple attempts.

In the case of identification, the enrolee reference is a potential candidate for the subject when (assuming that higher scores correspond to greater similarity) the comparison score exceeds a specified threshold, and/or when the comparison score is among the predetermined number of ranked values generated during comparisons across the entire database. The decision policy may allow or require multiple attempts before making an identification decision.

NOTE Conceptually, it is possible to treat multibiometric systems in the same manner as unibiometric systems, by treating the combined captured biometric samples, references or scores as if they were a single sample, reference or score and allowing the decision subsystem to operate score fusion or decision fusion as and if appropriate. (See also ISO/IEC TR 24722.)

6.2.7 Administration subsystem

The administration subsystem governs the overall policy, implementation, configuration and operation of the biometric system. Illustrative examples include:

a) interacting with the subject including providing guidance feedback to the subject during and/or after data capture, and requesting additional information from the subject;

b) storing and formatting of the biometric references and/or biometric interchange data;

c) providing final arbitration on output from decision and/or scores;

d) setting threshold values;

e) setting biometric system acquisition settings;

f) controlling the operational environment and non-biometric data storage;

g) providing appropriate safeguards for subject privacy and subject data security; and

h) interacting with the application that utilizes the biometric system.

NOTE The administration subsystem is not portrayed in Figure 1.

6.2.8 Interface to external application

The biometric system may or may not interface to an external application or system via a web services interface, an API, a hardware interface or a protocol interface.

NOTE The interface to external application is not portrayed in Figure 1.

6.3 Functions of general biometric system

6.3.1 Enrolment

In enrolment, a transaction by a capture subject is processed by the system in order to generate and store an enrolment reference for that individual.

Enrolment typically involves:

a) sample capture;

b) sample optimization or enhancement;

c) segmentation;

d) feature extraction;

e) quality checks (which may reject the sample/features as being unsuitable for creating a reference, and require capture of further samples);

f) presentation attack detection checks (which may reject the sample/features as being ineligible for use as an enrolment reference);

g) (where system policy so requires) comparison against existing biometric references to ensure the subject is not already enrolled;

h) reference creation (which may require features from multiple samples) and possibly generation of a database index;

i) storage of the biometric reference data record, possibly after conversion to a biometric reference data interchange format;

j) test verification or identification attempts by the capture subject to ensure that the resulting biometric reference is usable; and

k) allowing repeat enrolment attempts, should the initial enrolment be deemed unsatisfactory (dependent on the enrolment policy).

6.3.2 Verification of a positive biometric claim

In applications such as access control, a transaction by a subject may be processed by the system in order to verify a positive specific claim about the subject's enrolment (e.g. "I am enrolled as subject X"). Note that some biometric systems allow a single subject to enrol more than one instance of a biometric characteristic (for example, an iris system may allow subjects to enrol both iris images, while a fingerprint system may require enrolment of additional fingers for fallback in case a primary finger is damaged).

Verification of a specific positive claim typically involves:

a) sample capture;

b) sample optimization or enhancement;

c) segmentation;

d) feature extraction;

e) quality checks (which may reject the sample/features as being unsuitable for comparison, and require capture of further samples);

f) presentation attack detection checks (which may reject the sample/features as being ineligible for use)

g) probe creation (which may require features from multiple samples), possible conversion into a biometric data interchange format;

h) comparison of the probe and the reference for a biometric claim producing a comparison score;

i) determination of whether the biometric features of the probe match those of the reference based on whether the comparison score exceeds a threshold (in cases where higher scores correspond to greater similarity); and

j) decision to verify a claim based on the comparison result of one or more attempts as dictated by the decision policy.

The verification function either accepts or rejects the specific positive claim. The verification decision outcome is considered to be erroneous if either a false claim is accepted (false accept) or a true claim is rejected (false reject). In this application, a false acceptance occurs if the submitted sample is wrongly matched to a stored reference not created by the data subject. A false rejection occurs if the submitted sample is not matched to a reference actually created by the data subject.

NOTE Verification of an unspecific positive claim is also quite possible with a biometric system. Such applications have been called "PIN-less verification" because no PIN or other identifier was necessary to establish that the data subject was indeed enrolled in the database. The process is as above through steps a) – g). However, steps h) – j) are somewhat different when the claim is unspecific:

h) comparison of the probe against all the references producing a score for each comparison;

i) determination of whether the biometric features of the probe match those of any reference based on whether the comparison score exceeds a threshold (in cases where higher scores correspond to greater similarity); and

j) decision to verify a claim based on the comparison results of one or more attempts as dictated by the decision policy.

6.3.3 Identification

In identification, biometric samples from a capture subject are processed to generate a probe, and the enrolment database is searched to return identifiers of references similar to that probe. Identification provides a candidate list of identifiers containing zero, one, or more identifiers. Identification is considered correct when the subject is enrolled and an identifier for their enrolment is in the candidate list. The identification is considered to be erroneous if either an enrolled subject's identifier is not in the resulting candidate list (false-negative identification error), or if a transaction by a non-enrolled subject produces a non-empty candidate list (false-positive identification error).

Identification typically involves:

a) sample capture;

b) sample optimization or enhancement;

c) segmentation;

d) feature extraction;

e) quality checks (which may reject the sample/features as being unsuitable for comparison, and require capture of further samples);

f) presentation attack detection checks (which may reject the sample/features as being ineligible for use);

g) probe creation (which may require features from multiple samples) and possible conversion into a biometric data interchange format);

h) comparison against some or all references in the enrolment database, producing a score for each comparison;

i) determination of whether each compared reference is a potential candidate identifier for the capture subject, based on whether the comparison score exceeds a threshold and/or is among the highest ranked scores returned, producing a candidate list (higher scores correspond to greater similarity); and

j) an identification decision based on the candidate lists from one or more attempts, as dictated by the decision policy.

6.4 Enrolment, verification and identification transactions

The live acquisition processes for enrolment, verification or identification transactions consist of one or more capture attempts as allowed or required by the corresponding decision policy. Each capture attempt may consist of one or more presentations dependent on sensor operation, policy on sample quality, and any settings limiting the number of presentations or time permitted per attempt.

EXAMPLE 1 When a decision policy allows three attempts to verify, a transaction consists of one attempt, or two attempts if the first attempt is rejected, or three attempts if the first two attempts are rejected.

EXAMPLE 2 The enrolment process often requires the enrolee's biometric characteristics to be presented multiple times.

EXAMPLE 3 Some verification systems process a sequence of samples in a single attempt, for example: (a) collecting samples over a fixed period to find the best matching sample; or (b) collecting samples until either a match is obtained or the system times out.

EXAMPLE 4 Some biometric identification systems capture multiple samples from an individual and consolidate the results of biometric searches of each sample into a single candidate list. Typical cases are the use of multiple fingers in fingerprint identification, and the use of multiple frames in face recognition from video.

Figure 2 illustrates the relationship between presentations, attempts, and transactions.

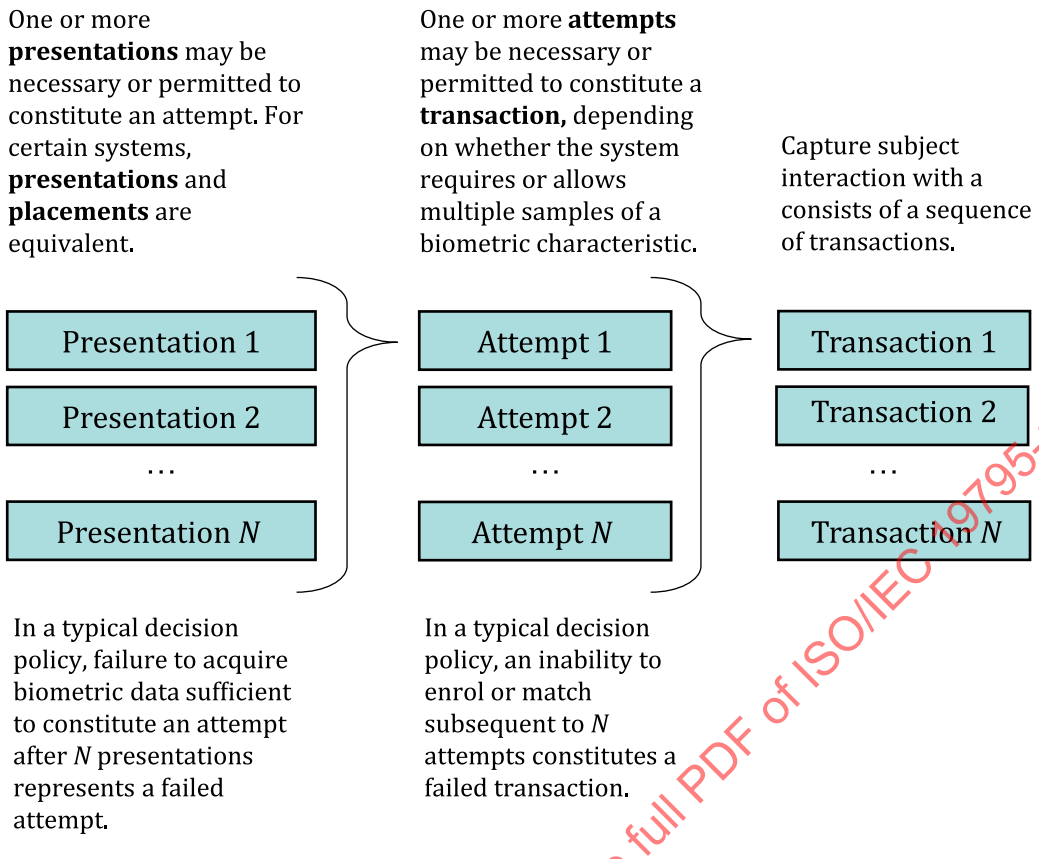


Figure 2 — Presentations, attempts and transactions

6.5 Performance measures

6.5.1 Error rates

In a biometric system, fundamental errors encompass comparison errors (false match and false non-match) and reference/probe creation errors (failure-to-capture, failure-to-extract, failure-to-enrol and failure-to-acquire). Fundamental errors combine to form decision errors for verification and identification systems. How these fundamental errors combine to form decision errors depends upon the number of comparisons required or allowed, whether there is a positive or negative biometric claim, and the decision policy.

6.5.2 Throughput rates

6.5.2.1 Throughput rates measure the number of subjects that can be processed per unit time based both on computational speed and human-machine interaction. Throughput rates for verification systems, such as those for access control, are usually determined by the speed of subject interaction with the system in the process of submitting a sufficient quality biometric sample. Throughput rates for identification systems, such as enrolment in a social service program, can also be affected by the computational processing time required to compare a captured sample to the biometric reference database. Therefore, depending on the application, it may be appropriate to measure the interaction times of subjects with the system and also the processing rate of the computational hardware as a function of the number of stored references. Further, in operational use, actions required subsequently to the biometric application decision (e.g. proceeding through a gate, opening a door, exiting a station) may affect the ability of the next subject to interact with the system. Measurement of throughput should take account of these additional actions which can vary depending on the biometric system decision.

NOTE Actual benchmark measurement of computer processing speed is covered in texts such as Reference [2] and is considered outside the scope of this document.

6.5.2.2 Measurement of the speed of subject interaction requires a precise definition of the actions indicating the initiation and termination of the transaction. This definition should be determined prior to the start of the test and noted in the test report. The test report should also include a brief listing of the actions of the capture subjects included in the transaction.

EXAMPLE It is possible to measure throughput for sample acquisition only, or for recognition transactions allowing multiple attempts.

6.5.3 Types of performance testing

6.5.3.1 Biometric technical performance testing can be of three types: technology, scenario or operational.

6.5.3.2 In a technology evaluation, testing of one or more algorithms is carried out on a corpus of biometric samples. The corpus may be collected as part of a test, or it may be a previously collected corpus available to the experimenter. Although example data may be distributed for developmental or tuning purposes prior to the test, the actual testing needs to be done on data that has not previously been seen by algorithm developers. Testing is carried out using offline processing of the data. As the corpus is fixed, the results of technology tests are repeatable. Nonetheless, performance against this corpus is dependent on both the environment and the population in which it is collected.

6.5.3.3 In a scenario evaluation, testing is carried out against one or more complete systems in an environment that models a real-world target application of interest. Each tested system has its own biometric capture device and so receives slightly different data. Consequently, if multiple systems are being compared, care is required to ensure that data collection across all tested systems is in the same environment with the same population. Depending on test requirements, testing can involve offline as well as online processing of data. Test results are repeatable only to the extent that the modelled scenario can be carefully controlled.

6.5.3.4 In an operational evaluation, testing is typically online for consistency with the application environment (unless the systems under test process data in an offline fashion). Repeatability of operational test results can be limited due to unknown and undocumented differences between operational environments. Furthermore, "ground truth" on the identities and behaviours of the biometric data subjects can be difficult to ascertain, particularly under unsupervised conditions without a test administrator, test observer or operational personnel present.

Annex A summarizes the different characteristics of the different types of evaluations.

7 Planning the evaluation

7.1 General

7.1.1 As the first step in an evaluation, the experimenter shall determine:

a) the systems / application / environment / population to be evaluated;

b) the aspects of performance to be measured; and

c) how a dataset for evaluating performance is to be created or identified (i.e. which is the appropriate evaluation type: technology, scenario, or operational?).

These decisions form the basis for developing an appropriate test protocol, specifying appropriate environmental controls, test subject selection and test size.

NOTE Sometimes the choice of evaluation type is determined by the availability of suitable databases of test samples for a technology evaluation, or of an installed system for an operational evaluation. Circumstances also exist in which all three types of testing are carried out sequentially, perhaps gradually narrowing down modality options and systems under consideration for the eventual deployment of a biometric system.

7.1.2 The test protocol may need to reflect system and application-specific differences such as:

a) differences in environments;

b) differences in capture subject populations (e.g. differences in subject habituation);

c) differences in biometric modalities (e.g. due to different modalities being affected by different environmental conditions, and the differences between testing of predominantly behavioural and predominantly physiological biometrics);

d) differences in the performance metrics of interest (e.g. overall performance of verification applications and identification applications are measured differently); and

e) additional problems in establishing the ground truth for identification systems (where identification transactions are not accompanied by a specific biometric claim).

7.1.3 This document provides the basic principles for conducting and reporting a performance evaluation. Other parts of the ISO/IEC 19795 series provide guidance and requirements for tests with specific evaluation types, biometric modalities and target applications.

7.2 Determine information about the system

The experimenter shall determine the following information about the system or systems to be tested in order to plan appropriate data collection procedures.

a) Does the system log transaction information? If not, then this information shall be recorded manually by the test subject, test administrator or test observer.

b) Does the system save biometric samples or features for each transaction? This is necessary if comparison scores based on data collected during the test are to be generated offline.

c) Does the system return comparison scores or just accept or reject decisions? In the latter case, data may have to be collected at a variety of threshold settings to generate a DET (see 8.2.3). If comparison scores are returned, what information is available regarding parameters and scale?

d) Is the vendor's SDK available? Offline generation of mated and non-mated comparison scores requires use of software modules from the SDK for:

   1) generating references from enrolment samples;

   2) generating probes from recognition samples; and

   3) determination of comparison scores between probes and references.

e) Are system modifications required for testing? Will required modifications alter system performance characteristics?

f) Does the system generate independent references? The correct procedures for collecting or generating non-mated comparison trials are different if references are dependent (see 8.4.11).

   NOTE An independent reference is one whose content and composition are not determined or informed by the content and composition of any other references.

g) Does the system use algorithms that adapt the reference after successful verification? In such cases, consideration shall be given as to how much reference adaptation should occur prior to measuring performance and also whether non-mated comparison trials are likely to adversely affect the references (see 8.4.5).

h) What are the recommended image quality and comparison score thresholds for the target application? These settings affect the quality of presented samples and error rates.

i) Are the expected approximate error rates known? This information can help in determining whether the test size is appropriate (see B.1).

j) What are the factors that influence performance for this type of system? These shall be controlled or documented (see 7.3).

k) Does performance depend on the number of enrolled references? This is the case for most identification systems, but also for some verification systems, such as those executing cohort enrolment, or those embedding a one-to-many search within the verification process.

7.3 Controlling factors that influence performance

7.3.1 Biometric system performance figures can be highly application-, environment- and population-dependent. Annex C provides a list of population, application, environmental and system factors that have been found to affect the performance of one or more types of biometric system.

7.3.2 Factors influencing the measured performance shall be explicitly or implicitly divided into one of four classes for control:

a) factors incorporated into the structure of the experiment (as independent variables) so that their possible effects can be observed, reported and analysed;

b) factors controlled to become part of the experimental conditions (unchanging throughout the evaluation);

c) factors randomized out of the experiment; or

d) factors judged to have negligible effect, which are ignored. Without this final category, the experiment becomes unnecessarily complex.

How these factors are to be controlled shall be decided in advance of data collection. This can involve some preliminary testing of systems to determine which factors are most significant and which may be safely ignored. In determining which factors to control, there can be a conflict between the needs for internal validity (i.e. differences in performance are due only to the independent variables recorded in the study) and external validity (i.e. the results truly represent performance on the target application).

EXAMPLE Suppose we are comparing the performance of two systems and are concerned over whether the skill or personality of the test administrator affects performance. Possible ways to control this factor are: a) to design the experiment to measure the performance differential between supervisors as well as between systems; b) to use only one supervisor, or to carefully script the supervisor/subject interaction to be as consistent as possible throughout the experiment; c) to allocate enrolment attempts randomly among all supervisors, thereby avoiding any systematic bias; or d) if there is prior evidence that differences between test administrators are small compared to the differences between systems, the experiment is permitted to ignore this factor.

7.3.3 For technology testing, a generic application and population may be envisioned, ensuring that the tests are neither too hard nor too easy for the algorithms being evaluated.

7.3.4 For scenario testing, a real-world application and population should be specified and modelled in order that the biometric system can be tested on representative subjects in a realistic environment.

7.3.5 In operational testing, the environment and the population are determined *in situ* with little control over them by the experimenter.

7.3.6 In scenario and operational testing, any adjustments to the devices and their environment for optimal performance (including quality and decision thresholds) should take place prior to data collection. Stricter quality control can result in fewer false matches and false non-matches, but a higher failure-to-acquire rate. The decision threshold also needs to be set appropriately if a comparison decision is presented to the capture subject — positive or negative feedback can affect subject behaviour. Vendors may advise on the optimal environment and trade-off between settings.

7.3.7 Of particular importance when planning the test is the time interval between enrolment and the collection of probe data. Longer time intervals generally make it more difficult to match probe samples to references due to the phenomenon known as "template ageing". This refers to the change in error rates caused by time-related changes in the biometric characteristic and its presentation. Collection of probe data for mated comparison trials shall therefore be separated in time from enrolment by an interval commensurate with the target application. If this interval is not known, then separation in time should be as long as is practicable.

NOTE Sometimes the elapsed time since enrolment is one of the controlled experimental factors, requiring collection of multiple samples over time. This enables a longitudinal analysis of the factors such as habituation, template ageing, and the stability/permanence of biometric characteristics over the timespan of the experiment. See 9.7.1.

7.4 Test subject selection

7.4.1 Both the enrolment and recognition functions require input signals or images. These samples should come originally from a test population or crew.

7.4.2 The test crew shall not include people whose biometric characteristics have previously been used to develop or tune the biometric system being tested.

7.4.3 The crew should be demographically similar to that of the target application for which performance will be predicted from test results. This is the case if test subjects are randomly selected from a target population. In other cases, volunteers can have to be relied upon.

7.4.4 Recruiting the crew from volunteers without imposing adequate controls can bias the tests. People with disabilities, for instance, could be under-represented in the sample population. It can be necessary to select from those volunteering to ensure that the test crew is as representative as possible and reflects the diversity of potential users. Target population approximation is an important consideration in ensuring the predictive value of tests.

7.4.5 Enrolment and recognition are normally carried out in different sessions, separated by days, weeks, months or years, depending on the target application. A test crew with stable membership over a long period is difficult to find and it should be expected that some test subjects drop out between sessions.

7.4.6 The test crew should be appropriately instructed so that their actions and behaviour follow that of the target application. If test subjects become bored or fatigued with routine testing, they can be tempted to experiment or be less careful. The experimenter shall consider test subject engagement as part of test design.

7.4.7 For applications where capture occurs without deliberate presentation, test subjects should ideally behave as if they were unaware of sample capture as it occurs. This may be achieved by passively capturing data over an extended period and by using RFID tags to establish test subjects' correct identifier without needing their input.

7.4.8 Prior to testing, test subjects should be fully informed about test duration, data collection procedures, how test data is used and disseminated, and how many sessions are required. Regardless of the use of the data, the identities of the crew should never be released by the experimenter. Test subject data management shall follow local regulations about data and privacy protection. These regulations may enforce, for example, the full disclosure of information and signature of consent form before taking part in an experiment and may also enforce that all acquired data are strictly anonymized.

NOTE For some types of testing, e.g. operational testing of covert identification systems, informing test subjects is impractical or presents a risk of altering their behaviour, thereby invalidating the collected results.

7.4.9 If it is necessary to use artificially generated samples or features (including those created by modifying real data) such use shall be reported and justified, and the method of generation and assumptions regarding appropriateness shall be described. Results for synthetic and non-synthetic data shall be reported separately, and results for mixed synthetic and non-synthetic data shall report details of the mixture.

NOTE The use of artificially generated images improves the internal validity of technology evaluations, as all the independent variables affecting performance are controlled. However, external validity is likely to be reduced. The database is also likely to be biased in respect of systems that model the biometric images in a similar way to that used in their generation.

7.5 Test size

7.5.1 General

The scale of an evaluation, in terms of the number of test subjects, instances and transactions, affects how accurately error rates are measured. The larger the test, the more accurate the results are likely to be. Rules such as the Rule of 3 and Rule of 30 (described in B.1) may be used to provide lower bounds to the number of attempts or transactions needed for a given level of accuracy. However, these rules overstate the confidence associated with biometric results. They assume that error rates are due to a single source of variability, which is not generally the case with biometrics.

NOTE As the test size increases, the variance of estimates decreases, but the scaling factor depends on the source of variability. For example, subjects can have differing error rates[3], giving a component of variance that scales as 1 / (number of test subjects) instead of 1 / (number of attempts). This effect is discussed in more detail in Annex B.

7.5.2 Collecting multiple recognition transactions per test subject per system

7.5.2.1 The evaluation may collect multiple transactions from each test subject per system. Circumstances in which several transactions should be collected from each subject per system include:

— testing the effects of ageing, habituation, and other systematic variations;

— testing of systems using reference updating;

— testing the extent to which different subjects have different individual error rates; or

— when the transaction is not fully defined prior to testing, e.g. to determine how varying the number of attempts per transaction alters performance.

NOTE If the cost and effort of obtaining and enrolling the test crew does not have to be considered, the ideal test has many test subjects, each making a single transaction. This provides independence between transactions. Ten probe-reference pairs from each of 100 people is not statistically equivalent to a single probe-reference from each of 1000 people and does not deliver the same level of certainty in the results. However, for cost and time efficiency, it is significantly easier to get existing enrolees to return than to find and enrol new test subjects. Furthermore, whenever a transaction is made, collecting additional transactions at the same time requires only marginally more effort. Despite the correlations between such multiple transactions, it is often the case that using multiple transactions from fewer test subjects produces a smaller uncertainty in test results than a trial of equal cost using a single transaction from slightly more test subjects.

7.5.2.2 The number and frequency of test transactions collected per test subject should be consistent with the target application. Use of multiple transactions per system would be inappropriate in tests where the capture subjects are expected to be unfamiliar with the device or biometric application.

NOTE Capture subject behaviour is likely to vary with each successive attempt due to increased familiarity with the device or feedback of the comparison decision. For example, the first attempt by a capture subject is possibly more likely to fail than their following attempts. As a result, the observed false reject rate depends on the pattern of attempts per subject, as defined by the test protocol. Generally, error rates are measured not only over the target population, but also over the types of attempt a subject reasonably makes. Averaging over multiple attempts helps in this case. However, there is the possibility that altering the number and pattern of attempts per subject affects subject behaviour enough to significantly alter the measured error rates.

7.5.3 Requirements on test size

The number of test subjects is more significant than the total number of test transactions in determining confidence level associated with test results.

a) The crew shall be as large as practicable. The measure of practicality is likely to be the expense of crew recruitment and management.

b) A given level of accuracy and confidence level required for the biometric performance test shall be determined as appropriate. Sufficient transactions shall be collected per test subject so that the total number of transactions exceeds that based on the level of accuracy and the confidence level determined. Rules such as the Rule of 3 and Rule of 30 may be used to determine the test size. If it is possible to collect transactions made by the same subject on different days, or with the subject using different instances (provided that the additional instances are representative of normal use), doing so can help reduce the dependencies between transactions by the same individual.

   NOTE 1 Use of the little finger is probably not representative of normal use of a fingerprint system, and the resulting error rates will be different[4]. Similarly, an inverted left hand is not representative in a right-handed hand geometry system.

c) Once data has been collected and analysed, the uncertainty in the performance measures shall be estimated.

   NOTE 2 The law of diminishing returns applies: a point will be reached where errors due to bias in the environment used, or in test subject selection, exceed those due to size of the crew and number of tests.

7.6 Multiple tests

7.6.1 The cost of data collection is such that it can be desirable to conduct multiple tests with one data collection effort. Technology evaluation allows for this. A single corpus may be collected for offline testing of multiple comparison algorithms. In effect, this decouples the data collection and signal processing subsystems. This is not problem-free however, as these subsystems are usually not completely independent. For instance, the quality control module, which can require the data collection subsystem to reacquire the image, is part of the signal processing subsystem. Furthermore, image quality can be affected by the device-specific user interfaces that guide the data collection process. Consequently, offline technical evaluation of algorithms using a common corpus might not reflect total system performance and might favour some systems over others.

7.6.2 Scenario evaluations of multiple systems can also be conducted over the same period by having a test crew use several different devices or scenarios in each session. However, this approach requires care. For example, as test subjects move from device to device, it is possible for habituation or acclimatization on one device to improve or worsen the presentation made on the subsequent device. To equalize such order effects over all devices, the order of their presentation to each test subject should be randomized.

8 Data collection

8.1 Avoidance of data collection errors

8.1.1 Collected biometric samples or features are properly referred to as a corpus. The information about those biometric samples or features and the subjects who produced them is referred to as the metadata. Both the corpus and the metadata can be corrupted by human error during the collection process. Errors attributable to mistakes in the collection process can easily exceed errors attributable to the biometric system. For this reason, extreme care shall be taken during data collection to avoid both corpus errors (wrongly acquired sample) and metadata errors (mislabelled sample).

EXAMPLE 1 Possible causes of corpus errors include: a) test subjects using the system incorrectly (and outside the limits allowed by the experimental controls), such as mistakenly using a fingerprint scanner upside down; and b) cases where a blank or corrupt image is acquired if the subject enters a PIN, but moves on before a proper image is captured.

EXAMPLE 2 Possible causes of metadata errors include: a) test subjects being issued with the wrong PIN; b) typing errors in PIN entry; and c) test subjects using the wrong body part, e.g. using a middle finger when the index finger is required.

8.1.2 Data collection software minimizing the amount of data requiring keyboard entry, multiple test administrators or test observers to double-check entered data and built-in data redundancy shall be used. Test administrators/observers shall be familiar with the correct operation of the system and the possible errors to guard against. To avoid a variable interpretation of what constitutes a wrongly acquired sample, objective criteria shall be set in advance. Any unusual circumstance surrounding the collection effort, and the transactions affected, shall be documented by test administrators/observers.

8.1.3 Even with precautions, some data collection errors are likely to be made, adding uncertainty to the measured test results. After-the-fact correction of corpus or metadata errors should be based on redundancies built into the collection system and should not be solely reliant on the outputs of the tested biometric algorithm. In this respect, systems that can save biometric samples and/or transaction logs offer more scope for error correction than systems where all the details have to be recorded manually.

8.1.4 Test administrators/observers shall not manually discard nor use an automated mechanism to discard samples collected for evaluation purposes unless the samples conform to a set of formal, pre-determined, documented and reported exclusion criteria. The number of samples so excluded shall be reported.

EXAMPLE Exclude fingerprint sample if captured area is less than 0,25 cm^2.

8.2 Data and details collected

8.2.1 The data that can be collected automatically depends on the biometric system implementation. Systems should automatically log all enrolment, verification or identification attempts, including details of claimed identity and comparison and quality scores. This minimizes the amount of data recorded by hand and the potential for transcription errors.

If systems save biometric samples, this brings the following advantages:

a) enrolment references and comparison scores can be generated offline, provided that a compatible SDK or API is available. This allows for a full cross-comparison of probes and references, giving a higher number of comparison scores;

b) the collected images may be reused to evaluate algorithm improvements or (provided the biometric samples are in a suitable format) to evaluate other algorithms in a technology evaluation; and

c) potential corpus or metadata errors may be checked by visually inspecting the images or through examining the transaction log.

8.2.2 Many biometric systems do not save samples or references, or automatically log all attempts, in their normal mode of operation. With vendor co-operation, it can be possible to incorporate this functionality into an otherwise standard system, but care should be taken that system performance is not affected. For example, the time taken in logging biometric samples or features can slow the system and affect subject behaviour. If biometric samples or features cannot be saved, enrolment, mated and non-mated, comparison trials shall be conducted online, and results recorded manually if necessary. This shall require closer supervision by the test administrators/observers to ensure that all results are logged correctly.

8.2.3 Some systems do not return comparison scores, but just a match or non-match decision at the current security setting. If the test protocol requires generation of a DET, results of mated and non-mated comparison trials shall be collected or generated at multiple threshold settings. The vendor may advise on the appropriate range of thresholds. The selected values for the threshold (which could be "low", "medium" and "high") may parameterize the DET in place of the decision threshold. In the case of online testing, each test subject shall execute the required number of transactions at each chosen security setting.

To ensure that the effects of multiple attempts are not confounded with those of changing the decision threshold, test protocols shall not allow mated comparison trials to be made at progressively more lenient settings, stopping when a match is obtained. Similarly, test protocols shall not allow non-mated comparison trials to be made at progressively stricter settings, stopping once a non-match is obtained.

8.2.4 The data collection plan should comply with applicable regulation(s) regarding data and privacy protection. The data collection plan may need to include a mechanism whereby a test subject may request their samples and biographical information to be expunged from the system. Otherwise, such redaction can be time-consuming and error-prone.

8.3 Enrolments

8.3.1 Enrolment transactions

8.3.1.1 Each test subject should enrol only once per system, though an enrolment may generate more than one reference within the enrolment record (for example a reference for each fingerprint, or multiple face poses). Multiple attempts may be allowed to achieve one good enrolment, with a predetermined maximum number of attempts or maximum elapsed time. Care shall be taken to prevent accidental multiple enrolments.

8.3.1.2 Sometimes it is appropriate to test performance of identification systems in which subjects can have multiple biometric reference data records. In such cases the number of test subjects and multiplicities of biometric reference data records shall be reported with the test results.

8.3.1.3 With some biometric modalities, a capture subject can present different biometric instances, e.g. up to ten fingers, left and right eye. To maximize the quantity of enrolment and recognition data available to the experimenter, an evaluation may treat enrolment of different instances from the same individual as separate enrolments. This practice shall be consistent with the intended mode of operations of the system under test. The experimenter shall document and report on such usage. Multiple enrolments from the same test subject in a given system shall not be considered equivalent to enrolment of different test subjects, e.g. for the purpose of achieving a level of statistical significance. See also 8.4.1.3.

8.3.1.4 Practice mated comparison trials may be performed immediately subsequent to enrolment to ensure that the enrolment samples are of sufficient quality to later match, and to familiarize test subjects with the system. Scores resulting from such practice trials should not be recorded as part of the mated comparison trial record.

8.3.2 Enrolment conditions

8.3.2.1 Enrolment conditions should model the target application enrolment. The taxonomy[5] of the enrolment environment determines the applicability of the test results. Vendor recommendations should be followed, and the details of the environment should be noted. Environmental conditions during enrolment require special attention. It is possible for conditions such as background noise in the case of speaker verification, or ambient lighting in the case of iris or face recognition systems, to reduce enrolment quality, impact enrolment error rates and impact resultant recognition rates. Of particular concern for systems with optical elements is light falling directly on the sensor, and uncontrolled reflections from the body characteristic being imaged. Lighting conditions should reflect the target application environment as closely as possible. Test results generated under one environmental condition often differ from those generated under different environmental conditions.

8.3.2.2 Every enrolment shall be carried out under the same general conditions. The goal should be to control presentation effects and channel effects so that such effects are either uniform across all test subjects, or randomly varying across test subjects. Many data collection efforts have been ruined because of changes in the protocols or equipment during the extended course of collection.

EXAMPLE A famous example is the "Great Divide" in the KING speech corpus[6]. About halfway through the collection, for a reason nobody now remembers, the recording equipment had to be temporarily disassembled. It was later reassembled according to the original wiring diagram; nonetheless the frequency response characteristics were slightly altered, creating a divide in the data and complicating the scientific analysis of algorithms based on the data.

8.3.2.3 As the tests progress, a test administrator can gain additional working knowledge of the system, which could affect the way later enrolments are carried out. The enrolment process and criteria for supervisor intervention shall be determined in advance and adequate supervisor training shall be provided.

8.3.3 Enrolment failures and presentation errors

8.3.3.1 The biometric system can reject some enrolment attempts on the basis of quality. For example, quality control modules for systems requiring multiple images for enrolment can reject transactions in which presentation quality varies significantly. Other quality control modules reject single poor-quality images. If these modules allow for tuning of the acceptance criteria, vendor advice should be followed. All enrolment quality scores should be recorded. Advice on remedial action to be taken with test subjects who fail an enrolment attempt shall be predetermined as part of the test plan.

NOTE Biometric sample quality is discussed in the ISO/IEC 29794 series (Parts 1, 4, 5 and 6).

8.3.3.2 The proportion of test subjects failing to enrol at the chosen criteria shall be recorded and reported as the failure-to-enrol rate. If possible, the reasons for enrolment failure should also be recorded and reported (e.g. those without the required biometric characteristics, cases where a sample could not be acquired, or failures/exceptions of the enrolment algorithm, or those unable to successfully verify in practice attempts).

8.3.3.3 Not all quality control is automatic. Intervention by the test administrator can be required if the enrolment biometric characteristic presented was erroneous according to some predetermined criteria. For instance, during enrolment test subjects can present the wrong finger, hand or eye, recite the wrong enrolment phrase or sign the wrong name. This data should be removed, but a record of such occurrences shall be kept.

8.3.3.4 Recognition testing subsequent to enrolment can reveal incorrectly captured enrolled references (for example, the wrong finger position was captured). Data editing to remove such enrolments can be necessary, but the effect of this on resulting performance shall be fully reported.

8.4 One-to-one comparison trials

8.4.1 General

8.4.1.1 The sampling plan shall ensure that the data collected are not dominated by a small group of excessively frequent, but unrepresentative subjects.

8.4.1.2 Mated comparison trials involve test subjects submitting samples to be compared against their own enrolment references. Non-mated comparison trials involve test subjects submitting samples to be compared against biometric references from different biometric data subjects.

8.4.1.3 In the cases where different instances from the same individual are being treated as separate enrolments (see 8.3.1.3), non-mated comparison trials shall not include intra-individual comparisons. Intra-individual comparisons are not equivalent to inter-individual comparisons. Moreover, in online use, it is all too easy for an individual to mistakenly present the wrong enrolled instance erroneously recording a false match.

EXAMPLE 1 The left index and middle fingerprints from a given test subject can share similarities that increase match rates relative to fingerprints from different people.

EXAMPLE 2 The vascular patterns of left and right palms from a given test subject have a different 'handedness' and are less similar than vascular patterns of the right palm from two different tests subjects.

8.4.1.4 The type of evaluation often determines whether an online or offline approach for comparison trials is used:

— In technology evaluation, comparison trials shall be conducted offline.

— In scenario evaluation, comparison trials shall be conducted online.

— In operational evaluation, comparison trials shall be conducted online.

— Sometimes the system tested in a scenario or operational evaluation is able to save probe and reference data. In such cases, if the system is also able to operate offline, or if the appropriate SDK is available, the saved data may be used to generate additional offline comparison trials reported alongside the online results.

8.4.2 Collection conditions

Comparison trial data shall be collected under environmental conditions that closely approximate those of the target application. This test environment shall be consistent throughout the collection process. The motivation of test subjects, and their level of training and familiarity with the system, should also mirror that of the target application. Non-mated comparison trials shall be made under the same conditions as mated comparison trials.

The collection process should ensure that presentation effects and channel effects are either uniform across all subjects or randomly varying across subjects. If the effects are held uniformly across subjects, then the same presentation and channel controls in place during enrolment should be in place for the collection of the test data. Systematic variation of presentation and channel effects between enrolment and test data can lead to results distorted by these factors. If the presentation and channel effects are allowed to vary randomly across test subjects, the experimenter should analyse results and report on any correlation in these effects between enrolment and comparison sessions.

8.4.3 Frequency of use

In the ideal case, between enrolment and the collection of comparison trial data, test subjects should use the system with the same frequency as the target application. However, this might not be a cost-

effective use of the test crew. It could be better to forego any interim use, but allow re-familiarization attempts immediately prior to comparison trials.

8.4.4 Systems performing optimization based on enrolled references

For systems that implement techniques such as score normalization based on the enrolled references to improve comparison performance, a sufficient number of test subjects should be enrolled prior to undertaking mated and non-mated comparison trials.

8.4.5 Systems performing reference adaptation

For systems that adapt the reference after successful verification, some interim use between enrolment and collection of mated comparison trial data can be appropriate. The amount of such interim use should be determined prior to data collection and should be reported with results.

Reference adaptation should be disabled during non-mated comparison trials. If this is not possible, non-mated comparison trials shall be delayed until all mated comparison trials are completed.

8.4.6 Processes for data entry errors and system misuse

Great care shall be taken to prevent data entry errors and to document any unusual circumstances surrounding the collection. Keystroke entry on the part of both test subjects and test administrators should be minimized. Data can be corrupted by test subjects who intentionally misuse the system. Every effort shall be made by test personnel to discourage these activities; however, data shall not be removed from the corpus unless external validation of the misuse of the system is available.

8.4.7 Failures to acquire

Capture subjects are sometimes unable to give a usable sample to the system as determined by either the test administrator or the quality control module. Test personnel should record information on failure-to-acquire attempts where these would otherwise not be logged. The failure-to-acquire rate measures the proportion of such attempts and is quality threshold dependent. As with enrolment, quality thresholds should be set in accordance with vendor advice.

NOTE Quality threshold (and decision threshold) settings can influence the behaviour of capture subjects: stricter thresholds encourage more careful presentation of the biometric pattern, looser thresholds allow more sloppiness. The database itself is therefore not necessarily threshold independent.

8.4.8 Adding test data to the corpus

All attempts, including failures-to-acquire, shall be recorded. Details shall be kept of the quality measures for each sample if available and, in the case of online testing, the comparison score or scores.

Transaction data shall be added to the corpus regardless of whether or not the comparison matches an enrolled reference. Some vendor software does not record transaction data unless the probe matches the enrolled reference. Data collection under such conditions would be severely biased in the direction of underestimating false non-match error rates. If this is the case, non-match errors shall be recorded by hand. Data shall be excluded only for predetermined causes, independently of comparison scores.

8.4.9 Online comparison trials

Online comparison trials are collected by having each test subject make recognition transactions:

a) against their own reference in the case of mated comparison trials; and

b) against each of a pre-determined number of non-self references randomly selected from all previous enrolments in the case of non-mated comparison trials. Selection of non-self references may be limited to enrolments within the same demographic group. The random selection shall be independent between data subjects. Because of the non-stationary statistical nature of the data

across data subjects, it is preferable that many test subjects conduct trials against a small number of randomly-chosen non-self references, as opposed to a small number of test subjects conducting trials against many non-self references.

Resulting comparison scores shall be recorded, together with the identifiers of both the capture subject and the reference. As mated and non-mated comparison trials can take place alongside each other, care shall be taken that results are attributed to the correct identity.

If a test subject is aware that a non-mated comparison is being made, changes in presentation behaviour can result in unrepresentative results, particularly with biometric systems that are based on predominantly behavioural characteristics. Therefore, to avoid even subconscious changes in presentation, test subjects should ideally not be told whether the current comparison is a mated or non-mated comparison trial.

Non-mated comparison trials may be conducted before all test subjects have enrolled. Though the first enrolled references have a higher probability of being targeted for a non-mated comparison, this does not bias the calculation of false match error rates if, as is usually the case, test subjects are enrolled in an order that has no regard to the quality of their biometric measures.

8.4.10 Offline comparison trials

Offline comparison trials are generated by comparing previously collected features against enrolment references. Previously collected features, or probes, may be from enrolled test subjects or from a distinct set of unenrolled test subjects. Offline computation allows a full cross-comparison approach in which each probe is compared against every reference.

The use of background databases of biometric samples or references acquired from different (possibly unknown) environments and populations is not considered best practice. Such usage shall be reported to include information on the background database size, characteristics, and other relevant factors.

Offline comparisons trials are made in the same basic way as online trials:

— For mated comparison trials comparing each probe with the corresponding mated reference;

— For non-mated comparison trials:

  — randomly selecting with replacement both samples and references for the non-mated comparisons trials;

  — randomly selecting, for each probe collected in a mated comparison trial, a number of non-self references from all those enrolled for comparison with the sample features (random selection of references being independent for each probe); or

  — performing a full cross-comparison, in which each collected probe is compared with every non-self reference. If there are N references and M probes (from the same test crew), $M(N - 1)$ non-mated comparison trials can be performed. These non-mated comparison trials are not statistically independent but, if each test subject makes the same number of transactions, this approach is statistically unbiased and represents a more efficient estimation technique than the use of randomly chosen non-mated comparisons[11].

Offline development of comparison scores should be carried out with software modules of the type available from vendors in SDKs. One module creates references from enrolment samples. A second module creates probes from recognition samples. Sometimes these modules are the same piece of code. A third module returns a comparison score for any assignment of a probe to a reference. Some vendors make offline evaluation tools available utilizing ground truth for each comparison to support the analysis of results. Such tools shall not be used other than for validation of the results generated by the test organization.

If an evaluation includes both online and offline comparison trials, the test administrator shall ensure that the same algorithm and the same parameter settings (e.g. the same background model) are used in both cases.

8.4.11 Offline non-mated comparison trials when references are dependent

For systems with dependent references, unbiased non-mated comparison trials may be generated using a jack-knife approach to create the enrolment references. The jack-knife approach is to enrol the entire crew with a single test subject omitted and a subset of non-mated comparison trials are generated between the omitted test subject and all the other enrolled references. This enrolment process is repeated omitting each crew member in turn, and a full set of non-mated comparison scores is generated.

A simpler technique may be used, in which the test crew is randomly partitioned into a set of enrolled and a set of un-enrolled test subjects. Mated comparison trials are based on recognition attempts by the enrolled test subjects, while non-mated comparison trials are based on recognition attempts by unenrolled test subjects. This is a less efficient use of the data than the jack-knife approach.

8.4.12 Offline non-mated comparison trials based on comparison of references

Cross-comparison of enrolment references may be used to generate non-mated comparison scores. This is useful, for example, in operational evaluations where samples or features of transactions are not saved. Each of N test (or enrolment) references may be compared to the remaining (N – 1) test (or enrolment) references. Reference cross-comparison shall not be used unless:

a) enrolment and verification are performed with the same capture subject interaction (for example, both require a single presentation);

b) enrolment and verification use the same sample feature extraction and encoding; and

c) quality thresholds for enrolment are the same as for verification attempts.

If these requirements are not fulfilled, reference cross-comparison is likely to result in biased estimation of non-mated comparison scores[5]. This is true whether the enrolment reference is averaged or selected from the best enrolment sample. No methods currently exist for correcting this bias.

8.4.13 Use of samples from multi-capture comparison transactions

Many biometric systems collect and process a sequence of samples in a single attempt, for example:

a) collecting samples over some fixed period, and scoring the best matching sample;

b) collecting samples until either a match is obtained or the system times out;

c) collecting samples until one of sufficient quality is obtained or the system times out; or

d) collecting a second sample when the score from the first sample is very close to the decision threshold.

In such cases, a single sample collected from a mated recognition attempt can be unsuitable as a probe for a non-mated comparison trial. For example, in case a), the collected sample is that which best matches the reference claimed in the mated comparison trial, and not the sample that best matches the reference of the non-mated comparison trial. To determine whether it is appropriate to base cross-comparison on data acquired for mated comparison trials, the following two questions shall be addressed:

— Does the saved sample depend on the reference being compared?

— If yes, does this materially affect the comparison scores generated?

If the answers to both these questions are yes, then either the whole sample sequence shall be saved and used in offline analysis, or non-mated comparison trials shall be performed online.

8.5 Identification trials

8.5.1 General

Identification trials shall be collected and recorded in the same general way as comparison trials, with the exception that the recorded outcome of identification trials shall consist of a candidate list containing zero, one, or more identifiers. It is strongly recommended that comparison scores for each candidate are returned to enable analysis at a range of operational thresholds. If quality scores are produced by the system these should also be recorded.

Identification searches may be conducted offline. Searches may be conducted against portions of the enrolment database of various sizes to record how identification performance varies with database size.

Identification may use pre-selection to limit the number of references compared by the comparison algorithm. In order to determine performance of the pre-selection algorithm, the number of pre-selected references for each identification attempt should be recorded (see Annex D).

8.5.2 Identification testing with non-enrolled test subjects

In addition to enrolled test subjects, identification testing shall include test subjects not enrolled in the system to ensure meaningful estimation of false-positive identification rate. These non-enrolled test subjects shall not be test subjects who failed enrolment. Identification transactions of enrolled capture subjects and of non-enrolled capture subjects should be made under the same conditions.

8.5.3 Use of jack-knife approach for identification testing

If the enrolment and identification samples of enrolled test subjects are stored, then unenrolled subject identification transactions may be generated offline using a jack-knife approach. In this case, the entire crew is enrolled with a single test subject omitted. The system then tries to identify the omitted test subject against the remainder of the test crew and the process is repeated for each test subject in turn. Considerations listed in 8.4.11 apply here as well.

9 Analyses

9.1 General

9.1.1 If the test crew is representative of the target population, and each test subject has one enrolment reference and makes the same number (and pattern) of transactions, the observed error rate proportions are the best estimates of the true error rates.

9.1.2 When the test crew is not representative of the target population (for example over-representation of known problem cases), or test transactions of individual test subjects are un-representative of those of the test crew as a whole (for example test subjects making more or fewer transactions than average), weighting may be appropriate to redress the imbalance. If error rates are estimated using a weighted proportion, the method of weighting shall be reported. When weighting by class of test subject is used, the observed class error rates should also be reported.

EXAMPLE If test subjects make differing numbers of verification or identification attempts, weighting of errors for each test subject in inverse proportion to the number of attempts the test subject makes avoids a potential bias towards the error rates of heavy users of the system.

9.1.3 It can be useful to measure error rates on a per-individual basis, per demographic (e.g. separate error rates for males and females), or per type of biometric instance (e.g. separate error rates for each finger position). Per-individual measures can be of intrinsic interest, indicating the type of person for whom better or worse performance is likely to be achieved. Per individual or per-demographic measures

are needed when the overall error rate is to be estimated as a weighted proportion. The extent of variation in per-individual error rates can help in estimating the uncertainty of performance estimates.

9.1.4 If errors in enrolment, sample acquisition and verification or identification are classified by cause, or by the step in the enrolment, acquisition or comparison process, then it can be possible to determine separate error rates for the different causes, or for the different components of the process.

9.2 Performance of biometric enrolment

9.2.1 Failure-to-enrol rate

9.2.1.1 The failure-to-enrol rate is the proportion of a specified set of enrolment transactions for which the system fails to create and store a biometric reference in accordance with the enrolment policy. The failure-to-enrol rate shall include:

— those unable to present the required biometric characteristic;

— those unable to produce a sample of sufficient quality at enrolment; and

— those who are unable to reliably produce a match decision with their newly created reference during attempts to confirm the enrolment is usable.

NOTE 1 The enrolment policy typically allows repeated enrolment attempts to obtain a reliable biometric reference.

NOTE 2 It is also possible to determine a failure-to-enrol rate for different biometric instances, such as for different fingers, for example, to report different failure-to-enrol rates for thumbs, index fingers, etc.

NOTE 3 In technology evaluations, analysis is based on a previously collected corpus and there is no problem in obtaining a biometric sample. Even so, enrolment failures sometimes occur. For example, when the biometric sample is of too low a quality for features to be extracted.

9.2.1.2 The failure-to-enrol rate for the target population shall be estimated as the proportion of enrolment transactions that failed under the predetermined enrolment policy.

9.2.1.3 The failure-to-enrol rate depends on the enrolment policy that governs the sample quality threshold for enrolment, the decision threshold to confirm the enrolment is usable, and the number of attempts or time allowed for enrolment in an enrolment transaction. The enrolment policy shall be described along with the observed failure-to-enrol rate.

NOTE Setting stricter quality requirements for enrolment increases the failure-to-enrol rate but improves performance of biometric comparison.

9.2.1.4 Attempts by test subjects unable to enrol in the system shall not be used for evaluation of acquisition performance or verification performance other than for the failure-to-enrol contribution to generalized false accept and false reject rates.

9.2.2 Enrolment transaction duration

The average enrolment transaction duration should be measured and reported.

In comparative scenario evaluations, it sometimes is not possible to set thresholds for time allowed or sample quality in a consistent manner across enrolment technologies. In these cases, one approach is to treat these variables as controlled factors of a given enrolment technology and measure how quickly sample capture is completed. It is then possible to compute the cumulative distribution of the failure-to-enrol rate as a function of enrolment duration (e.g. plotting FTER[duration] as a function of duration) for comparison across enrolment technologies.

9.3 Performance of biometric acquisition

9.3.1 Failure-to-acquire rate

9.3.1.1 The failure-to-acquire rate is the proportion of a specified set of acquisition processes that fail to produce biometric features suitable for biometric comparison. Failures to acquire include cases where:

— the capture subject is unable to present the biometric characteristics of interest (e.g. due to temporary illness or injury);

— the capture process fails;

— the segmentation or feature extraction processes fail;

— the extracted features fail to meet the quality control requirements of the system.

NOTE 1 It is also possible to determine a failure-to-acquire rate for transactions, e.g. measuring the proportion of transactions for which no attempts provided a sample of sufficient quality for comparison.

NOTE 2 In technology evaluations, the use of a previously collected test corpus precludes capture process failures. Nevertheless, failures to acquire are possible if samples are too low a quality for feature extraction.

9.3.1.2 The failure-to-acquire rate shall be estimated as the proportion of test attempts (for mated comparison trials) that were not able to be completed due to failures at presentation (no image captured), segmentation, feature extraction or quality control.

9.3.1.3 The failure-to-acquire rate depends on thresholds for sample quality, as well as the allowed duration for sample acquisition or allowed number of presentations. These settings shall be reported along with the observed failure-to-acquire rate.

NOTE Setting stricter quality requirements at enrolment increases the failure-to-enrol rate but improves matching performance.

9.3.1.4 Attempts where the sample was not acquired or did not meet quality thresholds are not processed by the comparison algorithm and do not generate comparison scores. Such failures-to-acquire shall be excluded in calculating the false match and false non-match error rates but shall be included in calculating the false accept and false reject rates. The failure-to-acquire, false match and false non-match rates shall be calculated at the same quality acceptance threshold settings.

9.3.2 Acquisition process duration

In comparative scenario evaluations, the acquisition process parameters that affect duration are unlikely to be consistent across technologies. In such cases, one approach is to treat these parameters as controlled factors for a given technology and measure how quickly sample acquisition is completed. The cumulative distribution of FTAR as the duration increases may then be calculated (e.g. plotting FTAR[duration] against duration) allowing fair comparisons between technologies.

9.3.3 Other aspects of acquisition performance

9.3.3.1 Sometimes the data collected during the acquisition process enables any failure to acquire to be attributed to failure of one of the subprocesses of acquisition. If failures are divided into failures of the capture process and failures of the feature extraction process, then a failure-to-capture rate (FTCR) and a failure-to-extract rate (FTXR) may be calculated and reported together with the combined failure-to-acquire rate (FTAR). In other cases, failure to acquire may be decomposed differently to include, for example, segmentation failure and failure to meet quality requirements, allowing for an alternative breakdown of the components of the failure-to-acquire rate to be reported.

9.3.3.2 With an appropriate test corpus providing a defined ground truth, it is possible to measure segmentation accuracy. The metric for segmentation accuracy necessarily varies depending on the biometric modality under consideration.

9.4 One-to-one comparison performance

9.4.1 False non-match rate

9.4.1.1 The false non-match rate is the proportion of completed mated comparison trials that result in a comparison decision of "non-match".

9.4.1.2 The false non-match rate depends on the comparison score threshold and shall be quoted along with the observed false match rate at the same threshold (or shown against the false match rate at the same threshold in a DET plot).

9.4.1.3 In evaluations where test subjects have made multiple attempts, it can be useful to show how the false non-match rate varies over the test crew. This may be done by calculating an error rate for each test subject's attempts, and plotting a histogram showing the error rate for each test subject, ordering test subjects in increasing order of their error rates.

9.4.2 False match rate

9.4.2.1 The false match rate is the proportion of a specified set of completed non-mated comparison trials that result in a comparison decision of "match".

NOTE In non-mated comparison trials, probes are normally acquired from capture subjects presenting their personal biometric characteristic as if they were attempting successful verification against their own reference, and without any attempt at impersonation. For example, in the case of dynamic signature verification, the capture subject signs their own name rather than a name corresponding to the non-mated reference. In such cases, where aspects of the required biometric characteristic are easily imitated, a further set on non-mated comparison trials can be collected in which a degree of imitation is allowed, producing a separate false match rate for that level of impersonation. However, defining the methods or level of skill to be used in impersonations is outside the scope of this document.

9.4.2.2 The false match rate depends on the comparison score threshold and should be quoted along with the observed false non-match rate at the same threshold (or shown against the false non-match rate at the same threshold in a DET plot).

9.4.2.3 If a test subject is enrolled, and their enrolment affects the references of others in the system, or if the comparison algorithm modifies itself using this (and other) references, then non-mated comparison trials using that subject are biased and should not be used to estimate the false match rate. Subclause 8.4.11 details how to deal with such cases.

EXAMPLE Eigenface systems, using all enrolled images for creation of the basis-images, and cohort-based speaker recognition systems are two examples for which references are dependent.

9.4.2.4 Comparison of genetically identical biometric characteristics (for instance, between an individual's index and middle fingers, or across identical twins) yields different score distributions than comparison of genetically different characteristics[12]-[14]. The test plan should declare the policy on whether non-mated comparisons known to be between identical twins, siblings, or parent and child are to be included in deriving the false match rate. Comparisons between two biometric instances from the same individual should always be excluded.

NOTE 1 The incidence of identical (monozygotic) twins is approximately 3 or 4 in 1000 births and it is possible that their presence in a deployed face recognition system increases the false match rate above that measured excluding such genetically similar comparisons.

NOTE 2 Genetic effects extend beyond just identical twins into families and national populations. In face recognition for example, this sometimes results in much higher false match rates being recorded between subjects born in the same country than between subjects from geographically and ethnically distinct countries.

9.4.2.5 In evaluations where there are several non-mated comparison trials per test subject, or per reference, it is useful to show how the false match rate varies over test subjects and over stored references. This involves calculating the individual false match error rate for the biometric reference of each test subject and for the biometric probes from each test subject. Histograms may be plotted to show the error rate for each test subject, ordering test subjects in increasing order of their error rates.

EXAMPLE It is possible for a face recognition system to admit a set of "golden faces" where false matches occur mainly with this set of faces. Histograms showing variation of error rates across subjects reveal this vulnerability.

9.5 Verification system performance metrics

9.5.1 General

A first order estimation of the false accept and false reject rates for transactions of multiple attempts can be derived from the detection error trade-off curve. However, such estimates are unable to take account of correlations in sequential attempts and in the comparisons involving the same subject and consequently are sometimes quite inaccurate. Therefore, these performance metrics shall be derived directly, using test transactions with multiple attempts as specified by the decision policy.

9.5.2 False reject rate

9.5.2.1 The false reject rate is the proportion of a specified set of verification transactions with true biometric claims erroneously rejected. A transaction may consist of one or more attempts depending on the decision policy.

9.5.2.2 Verification transactions rejected due to failures-to-acquire are included in the count of false rejections along with those denied due to comparison errors. In the case of verification systems with positive biometric claims (e.g. for access control), acceptance requires both successful acquisition of the biometric characteristics and a comparison decision of "match".

EXAMPLE If, for a verification system with positive biometric claims, a transaction consists of a single attempt, then a failure–to-acquire or a false non-match causes a false rejection, and the false reject rate is given by:

$$FRR = FTAR + FNMR(1 - FTAR)$$

where:

FRR is the false reject rate;

FTAR is the failure-to-acquire rate;

FNMR is the false non-match rate.

9.5.2.3 The false reject rate depends on the decision policy, the comparison score threshold, any threshold for sample quality as well as the allowed duration or allowed number of presentations. The false reject rate shall be reported with these details, alongside the estimated false accept rate at the same values (or plotted against the false accept rate at the same threshold(s) in a DET plot).

9.5.3 False accept rate

9.5.3.1 The false accept rate is the proportion of a specified set of transactions with false biometric claims erroneously accepted. A transaction may consist of one or more attempts depending on the decision policy.

9.5.3.2 Rejections due to failures-to-acquire are counted in the number of transactions, but not in the number of false acceptances. In the case of verification systems with positive biometric claims (e.g. for access control), acceptance requires both successful acquisition of the biometric characteristics and a comparison decision of "match".

EXAMPLE If, for a verification transaction with positive biometric claims, a transaction consists of a single attempt, then a successful acquisition followed by a false match results in a false acceptance and the false accept rate is given by:

$$FAR = FMR\ (1 - FTAR)$$

where:

FAR is the false accept rate;

FMR is the false match rate;

FTAR is the failure-to-acquire rate.

9.5.3.3 The false accept rate depends on the decision policy, the comparison score threshold, and any threshold for sample quality, as well as the allowed duration or allowed number of presentations. The false accept rate shall be reported with these details, alongside the estimated false reject rate at the same values (or plotted against the false reject rate at the same threshold(s) in a DET plot).

9.5.4 Verification transaction duration

9.5.4.1 In comparative scenario evaluations, the acquisition process parameters that affect duration are unlikely to be consistent across technologies. In such cases, one approach is to treat these parameters as controlled factors of a given technology and measure how quickly a verification transaction is completed. The cumulative distribution of FRR as duration decreases may then be calculated (e.g. plotting FRR[duration] against duration) allowing fair comparisons between technologies.

9.5.4.2 To reduce the effect of outlier long transaction durations due to poor habituation or interruptions during the evaluation process, the average duration for a successful verification transaction shall be calculated in the following manner.

a) For each test subject, the median duration of successful mated verification transactions is determined;

b) These values are then averaged over all (enrolled) test subjects.

9.5.5 Generalized false reject rate and generalized false accept rate

Comparison of systems with different failure-to-enrol rates can require use of generalized false reject and false accept rates which combine failure rates of the enrolment and acquisition processes with error rates of the comparison process. The method of generalization should be appropriate to the evaluation. A typical generalization is to treat a failure-to-enrol as if the enrolment was completed but

all subsequent verification transactions by that enrolee, or against their reference are treated as a "non-match". The method of generalization shall be reported.

EXAMPLE 1 We assume a scenario evaluation for a verification system with positive biometric claims allowing a single attempt per verification transaction. Test subjects who are not enrolled take no further part in the evaluation. The test protocol is to conduct a full cross-comparison between probes and references of all test subjects. In this case, a generalized false acceptance occurs when (i) the subject providing the probe and the subject providing the reference were both enrolled, (ii) there is successful acquisition of the probe, and (iii) there is a false match. A generalized false rejection occurs if (i) the subject is not enrolled, or (ii) the submitted sample cannot be acquired, or (iii) there is a false non-match. In this example, the generalized false accept and false reject rates are given by:

$$GFAR = FMR\ (1 - FTAR)\ (1 - FTER)^2$$

$$GFRR = FTER + (1 - FTER)\ FTAR + (1 - FTER)\ (1 - FTAR)\ FNMR$$

where:

GFAR is the generalized false accept rate;

GFRR is the generalized false reject rate;

FMR is the false match rate;

FNMR is the false non-match rate;

FTER is the failure-to-enrol rate;

FTAR is the failure-to-acquire rate.

EXAMPLE 2 In a technology evaluation, enrolment references are generated from all reference samples that do not cause a failure-to-enrol, and attempt features are generated from all probe samples that do not cause a failure-to-acquire. In this case, the generalized false accept and false reject rates are given by:

$$GFAR = FMR\ (1 - FTAR)\ (1 - FTER)$$

$$GFRR = FTER + (1 - FTER)\ FTAR + (1 - FTER)\ (1 - FTAR)\ FNMR$$

9.6 Identification system performance metrics

9.6.1 General

Identification system error rate metrics are dependent on the operational characteristics and settings of the comparison system/algorithm being used. There are three primary parameters defining these characteristics:

— *N*, the number of references in the enrolled dataset;

— *R*, the rank index, where only candidates at ranks 1 through *R* are considered as potential identifiers for the capture subject (*R* is typically bounded by a value *L*, the length of the candidate list returned by the identification system tested); and

— *T*, the comparison score threshold used to determine whether a candidate is a potential match for the capture subject.

In order to precisely compute and report the identification error rate metrics, the values of these characteristics shall be declared.

NOTE If *R* is set equal to *N*, then candidates are returned solely on the basis of candidate score exceeding threshold *T*. Similarly, if *T* is set to 0 (assumed to be the minimum possible similarity score) then candidates are returned solely on the basis of candidate rank.

9.6.2 False-negative identification rate

9.6.2.1 The false-negative identification rate is the proportion of a specified set of identification transactions by capture subjects enrolled in the system (mated identification transactions), for which the subject's correct identifier is not included in the candidate list returned.

$$\mathrm{FNIR} = \frac{\left|\left\{i \in M_D \,|\, m_i \notin C_i\right\}\right|}{\left|M_D\right|}$$

where:

M_D is the set of mated identification transactions with reference database *D*;

m_i is the mated reference for transaction *i*;

C_i is the candidate list for identification transaction *i*; and

$|.|$ denotes the cardinality (number of elements) of set.

The candidate list comprises identifiers for the biometric references which are sufficiently similar to the identification transaction probe. Sufficient similarity is generally based on the candidate rank being in the range 1 to *R*, or the candidate score exceeding the threshold, *T*. Then, the false-negative identification rate is the proportion of non-mated identification transactions in which either the mated reference has rank greater than *R* or the mated reference has a candidate score below *T*.

$$\mathrm{FNIR}(N,R,T) = \frac{\left|\left\{i \in M_D \,|\, \left(rank_i\left(m_i\right) > R\right) \mathrm{or} \left(score_i\left(m_i\right) \leq T\right)\right\}\right|}{\left|M_D\right|}$$

where:

M_D is the set of mated identification transactions with reference database R;

m_i is the mated reference for transaction *i*;

$rank_i(\,)$ gives the candidate rank of a reference in identification transaction *i*; and

$score_i(\,)$ gives the candidate score of a reference in identification transaction *i*.

9.6.2.2 The experimenter may compute FNIR(*N, R, T*) for *N*, *R* and *T* in their appropriate ranges.

9.6.3 False-positive identification rate

9.6.3.1 The false-positive identification rate is the proportion of a specified set of identification transactions by capture subjects not enrolled in the system (non-mated identification transactions), where a reference identifier is returned.

$$\mathrm{FPIR} = \frac{\left|\left\{i \in U_D \,|\, \left|C_i\right| > 0\right\}\right|}{\left|U_D\right|}$$

where:

U_D is the set of non-mated identification transactions with reference database *D*; and

C_i is the candidate list for identification transaction *i*.

For an identification transaction to return a candidate reference identifier, the top-ranked candidate must have a candidate score exceeding the threshold. Thus, the false-positive identification rate may

also be stated as the proportion of non-mated identification transactions for which the candidate score of the top-ranked reference exceeds the specified threshold.

$$\mathrm{FPIR}(N,T)=\frac{\left|\left\{i\in U_D \,|\, score_i\left(t_i\right)>T\right\}\right|}{\left|U_D\right|}$$

where:

U_D is the set of non-mated identification transactions with reference database R;

t_i is the top-ranked reference identifier in identification transaction *i*; and

$score_i(\,)$ gives the candidate score of a reference in identification transaction *i*.

The relationship between the false-positive identification rate FPIR(*N*, *T*) and false-negative identification rate at rank one FNIR(*N*, 1, *T*) as the threshold *T* varies may be plotted as a DET.

9.6.3.2 As the enrolment database of an identification system grows, the change in overall identification performance for a constant value of the false-positive identification rate (requiring the threshold T_N to be adjusted with database size) may be shown as a plot of the false-negative identification rate FNIR(*N*, 1, T_N) against size of enrolment database *N*. Alternatively, a set of DET plots showing the relationship between false-positive and false-negative identification rates may be plotted for various database sizes, as in the example shown in Figure E.1.

9.6.4 Generalized false-negative identification rate and generalized false-positive identification rate

Comparison of identification systems having different failure-to-enrol rates and failure-to-acquire rates requires the use of generalized false-negative and generalized false-positive identification rates that combine failures to enrol, failures to acquire and recognition errors. The method of generalization should be appropriate for the evaluation. The method of generalization shall be reported.

NOTE Typical generalizations are to treat a failure to enrol as if the enrolment was completed, but the reference identifier does not appear in any candidate list, and to treat a failure to acquire as if the search was completed and no candidate was found.

9.6.5 Selectivity

In identification systems configured to produce multiple candidate identities, the quantity selectivity is the average number of candidates returned for which the similarity to the probe exceeds a threshold, *T*, in a non-mated identification transaction:

$$\mathrm{SEL}(N,R,T)=\frac{1}{|U|}\sum_{i\in U}\left|C_i\left(R,T\right)\right|$$

where:

U is the set of non-mated identification transactions conducted; and

$C_i\left(R,T\right)$ is the candidate list returned for identification transaction *i* with candidate rank no greater than *R* and candidate score greater than *T*.

NOTE 1 Selectivity is sometimes measured against the entire database. This would be SEL(*N*, *N*, *T*).

NOTE 2 Selectivity and FPIR differ at lower thresholds, converging as false positives become rare at higher thresholds. However, note that false positives are sometimes concentrated in certain transactions.

9.6.6 Closed-set test of identification performance

Identification performance may be shown as a CMC plot derived from a closed-set test. The CMC plots the true positive identification rate TPIR(N, R, 0) as a function of R (see 10.4).

The true positive identification TPIR(N, R, 0) is the proportion of identification transactions by a subject enrolled in the system for which the subject's reference identifier is within the top R candidates returned. TPIR(N, R, 0) = 1 – FNIR(N, R, 0).

While use of the CMC is very common, if such plots are unable to adequately display the range of identification rates (e.g. in comparative tests where values can readily span more than one order of magnitude) the experimenter should instead plot FNIR(N, R, 0) against R on a logarithmic scale.

NOTE A suggested algorithm for generating data points on the CMC plot is provided in Annex F.

9.6.7 Estimation of identification error rates from one-to-one comparison results

A first order estimation of the false-positive and false-negative identification rates for identification systems may be derived from knowledge of the FMR-FNMR detection error trade-off. However, such estimates ignore possible correlations between separate comparisons involving the same subject, and consequently there is a risk of inaccuracy.

EXAMPLE The performance of an identification algorithm using a single biometric sample against a reference database of size N can be approximated using the following formulae, providing these formulae have been validated by the identification error rates observed on the test data.

$$\mathrm{FNIR}(N,1,T) = \mathrm{FTAR} + (1 - \mathrm{FTAR})\,\mathrm{FNMR}(T)$$

$$\mathrm{FPIR}(N,T) = (1 - \mathrm{FTAR})\,(1-(1-\mathrm{FMR}(T))^{N})$$

where

FPIR(N,T) is the false-positive identification rate at threshold T;

FNIR (N,1,T) is the false-negative identification rate at rank 1 and threshold T;

FTAR is the failure-to-acquire rate;

FMR (T) is the false match rate at threshold T;

FNMR (T) is the false non-match rate at threshold T;

N is the number of references in the database; and

T is the comparison score threshold.

NOTE 1 In the case of identification systems using pre-selection, the above model of performance can be extended using the performance metrics for the pre-selection algorithm (see Annex D).

NOTE 2 In some cases, FPIR does not scale according to the binomial formula because the search is not implemented as N independent 1:1 comparisons. This is the case if a fast tree-based search is used, or if scores are computed with a dependence on other references. In such cases, as N increases FPIR remains fixed but FNIR is expected to increase as similar non-mated references displace correct mates.

9.6.8 Predicting identification error rates in larger populations

Estimation of the performance of large-scale identification systems (beyond the size of the test) may need to be extrapolated using both the first-order estimation and the identification performance on the smaller database. The model used for extrapolating performance shall be reported in such cases.

9.7 Analysis of performance across controlled experimental factors

9.7.1 Longitudinal analyses

9.7.1.1 Of particular concern to the long-term use of a biometric reference is whether the corresponding biometric characteristic changes irreversibly due to ageing. The term "stability" (or "permanence"[15]) requires that a biometric characteristic should be sufficiently invariant over time with respect to a given comparison algorithm. A lack of stability is the component of template ageing that is intrinsic to the biometric source and is essentially unavoidable and irreversible.

EXAMPLE In face recognition, appearance changes as soft tissues lose their elasticity, muscle tone and volume.

9.7.1.2 Ageing manifests as a change to mated comparison scores. Longitudinal analysis of scores produced by comparing samples collected over time can yield statements of permanence. These can be in the form of trends in comparison scores and increases in FNMR. Mixed effects regression models are recommended because they are capable of handling imbalance across subjects, irregular sampling, shared population effects and random individual-specific effects. For longitudinal analysis, see [16] and for its application in biometrics see [17] and [18].

9.7.1.3 For some biometrics, accuracy depends on age even with a fixed time lapse between samples. Procedures for analysing both age and ageing have been developed[17].

9.7.2 Pairwise analyses

In cases where subjects each use two different devices or participate under two different experimental conditions the data may be subject to pairwise tests or plotting of the difference in outcome (e.g. showing differences in score distribution or transaction duration).

9.8 Detection error trade-off

9.8.1 The detection error trade-off (DET) shall be developed using the comparison scores from the mated and non-mated comparison trials. The mated and non-mated comparison scores are to be ordered. Any outliers should be investigated to determine if labelling errors are indicated. Removal of any scores from the test should be fully documented and will lead to external criticism of the test results.

9.8.2 The DET is established through the accumulation of the ordered mated and non-mated comparison scores. As the score varies over all possible values, each point (x, y) of the DET is derived from the false match and the false non-match rate using that score as the decision threshold. Assuming higher comparison scores show greater similarity between probe and reference, the false match rate is the proportion of non-mated comparison scores at or above (more similar) the current value of the score parameter, and the false non-match rate is the proportion of mated comparison scores below (less similar) the score parameter. The DET may be tabulated as a list of triplets (FMR, FNMR, threshold), or shown as a DET plot (see 10.2).

NOTE A suggested procedure for deriving the data points on the DET is provided in Annex F.

9.8.3 The DET can also be used to show the relationship between the false accept rate and false reject rate in a similar manner. The false accept rate and false reject rate relate to the false match rate, false non-match rate, and failure-to-acquire rate in a manner that depends on the number of attempts allowed by the decision policy. Transactions of multiple attempts can require generation of a new transaction score based on the similarity scores of the constituent attempts (e.g. the maximum value of the similarity scores for a best of three attempts decision policy). Similarly, DET plots may be used to show the relationship between identification error rates.

9.9 Transaction durations

9.9.1 Throughput rates measure the number of subjects that can be processed per unit time based both on computational speed and human-machine interaction. Throughput rates for verification systems, such as those for access control, are usually determined by the speed of subject interaction with the system in the process of submitting a sufficient quality biometric sample. Throughput rates for identification systems, such as enrolment in a social service program, can also be affected by the computational processing time required to compare a captured sample to the database of stored references. Therefore, depending on the application, it can be appropriate to measure the interaction times of subjects with the system and also the processing rate of the computational hardware as a function of the number of stored references.

9.9.2 Measurement of the speed of subject interaction requires a precise definition of the actions indicating the initiation and termination of the transaction. This definition should be determined prior to the start of the test and noted in the test report. The test report should also include a brief listing of the actions of the capture subjects included in the transaction. In operational use, actions required subsequent to the biometric result (e.g. proceeding through a gate, opening a door, exiting a station) can affect the ability of the next subject to interact with the system. Measurement of throughput should take account of these additional actions which can vary depending on the biometric system result.

9.9.3 Recognizing that transaction durations can extend far beyond the average, the cumulative distribution function of transaction times should be shown when reporting transaction durations.

9.10 Computational workload

9.10.1 Some biometric applications, such as identification searches against large enrolment databases, or generation of biometric probes from video footage, can be computationally intensive, and it is useful to measure computational workload. As such applications scale, hardware resources available can limit throughput or accuracy. Measures of computational workload include transaction time and memory usage and may extend to consideration of aspects such as CPU usage, and network and disk activity.

NOTE 1 Computational workload of biometric search can depend on the number of enrolled references and the size of the references.

NOTE 2 Computational workload can be measured separately for different elements of a biometric transaction such as pre-selection.

9.10.2 In order to report the total computational requirements of a single transaction (enrolment, verification, identification) in a biometric system, the computational workload shall be measured over all components of the transaction:

a) Enrolment:

   — Generation of a biometric enrolment data record.

   — Generation of a biometric index, if any.

   — Duplicate enrolment check (which corresponds to an identification search against existing enrolment references) if it is implemented.

   — Storage in the reference database.

b) Verification:

   — Generation of a biometric feature set from the captured biometric sample.

   — Retrieval of the biometric enrolment record corresponding to the claimed identity.

   — Comparison-trial, including additional costs, e.g. alignment (such as the bit-shifts in iris-codes).

— Accept/reject decision.

c) Identification:

— Generation of a biometric feature set from the captured biometric sample.

— Pre-selection to reduce workload of identification search if it implemented.

— Identification search over the reference database.

— Production of candidate list and deciding identification outcome.

NOTE The computational workload of identification generally increases as the number of enrolled references increases. It is possible for the computational workload for an unmated biometric probe to be higher than that for a mated biometric probe where the mated reference is found before the entire enrolment database is searched.

9.10.3 For comparison of the performance of different biometric algorithms or systems, it is necessary to take account of the different computational power of the hardware on which the systems are operating; for example, using relative performance against a common benchmark algorithm. If possible, computational workload of the compared systems or algorithms should be measured using the same hardware and configuration.

9.10.4 For comparison of biometric identification systems which are intended to maintain the recognition accuracy and to decrease computational workload (e.g. through use of binning, preselection, or indexing algorithms) the test report shall tabulate or plot results for each system showing:

— computational workload, and

— recognition accuracy (FPIR, FNIR)

over the range of number of enrolled references.

Metrics applicable to the workload reduction method should be reported. For example, in the case of pre-selection, the preselection error rate and penetration rate should be reported. Any model used to extrapolate computational workload as the number of enrolled references increase should also be reported.

9.11 Uncertainty of estimates

9.11.1 Performance estimates are affected by both systematic errors and random errors. Random errors include those due to the natural variation in test subjects and sample presentation. Systematic errors include those due to bias in the test procedures, for example if certain types of individual are under-represented in the test crew. Neither type of error is perfectly quantifiable and therefore there will be an uncertainty in the results of a performance evaluation. Nevertheless, the uncertainty in the measured performance shall be estimated. Annex B provides some methods by which random uncertainty in performance results may be estimated.

9.11.2 Uncertainties arising from random effects become smaller as the size of the test increases and can often be estimated from the collected data. Systematic uncertainty persists regardless of test size. It may be possible to determine the effects of some of the systematic errors. For example, checking whether the error rates for an under-represented category of individuals are consistent with the overall error rates could show whether a properly balanced test crew would give different error rates. Part of the performance trial may be repeated in different environmental conditions to check that the measured error rates are not unduly sensitive to small environmental changes.

10 Graphical presentation of results

10.1 Score distributions

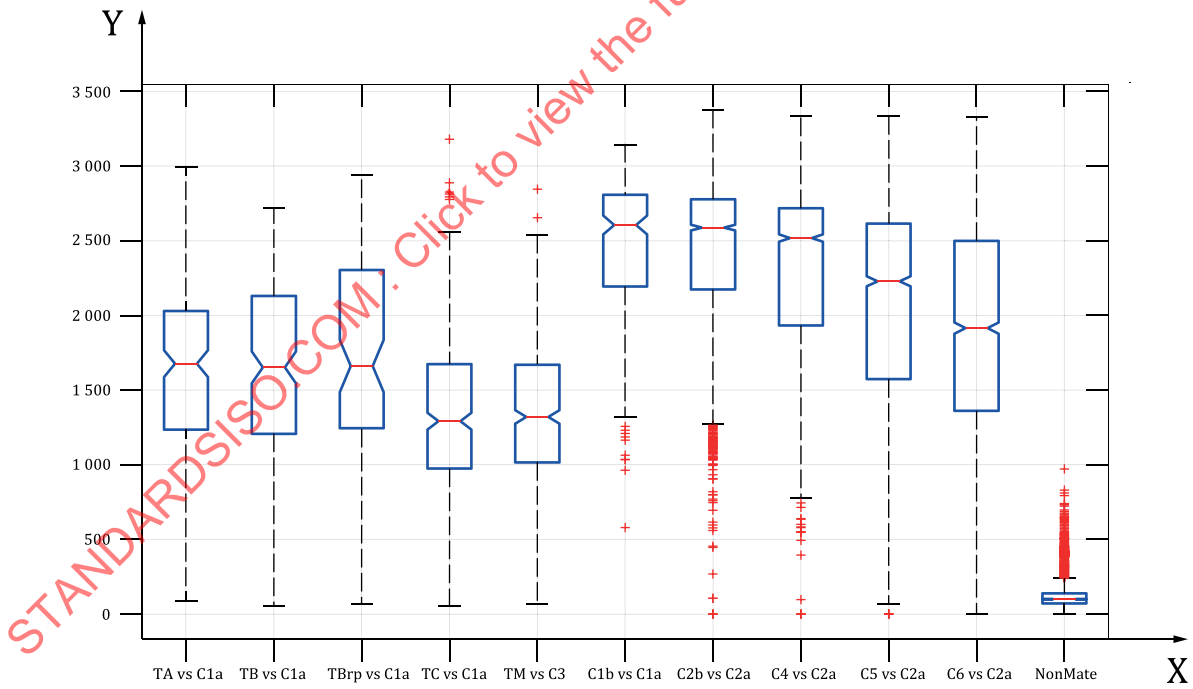
10.1.1 General

Histograms for both mated and non-mated comparison scores can be instructive, particularly when illustrating the variation in comparison scores for transactions made under differing conditions. Due to the range and spread of proprietary comparison score distributions, prior to plotting score histograms, rescaling of the scores should be considered to aid comparison of results within and between systems.

10.1.2 Boxplots

Another method for visual summary of score distributions is known as the boxplot or box and whisker plot. This method enables simultaneous display of measurement distributions for multiple experimental conditions, or devices or comparisons of measurements between devices.

Figure 3 shows an example boxplot from NIST SP 500-305, comparing the distribution of mated comparison scores for various touchless-to-contact, and contact-to-contact fingerprint recognition systems. The figure uses "notched boxplots". The median values are surrounded by a notch which is sized such that overlapping notches imply that the differences in median are not statistically significant. The boxes represent the interquartile range of each set of score data. Datapoints which are more than 1,5 times the interquartile range above the upper quartile, or below the lower quartile, are classed as outliers and shown as a cross (+). The whiskers extend from the box to the highest and lowest scores that are not classed as outliers.



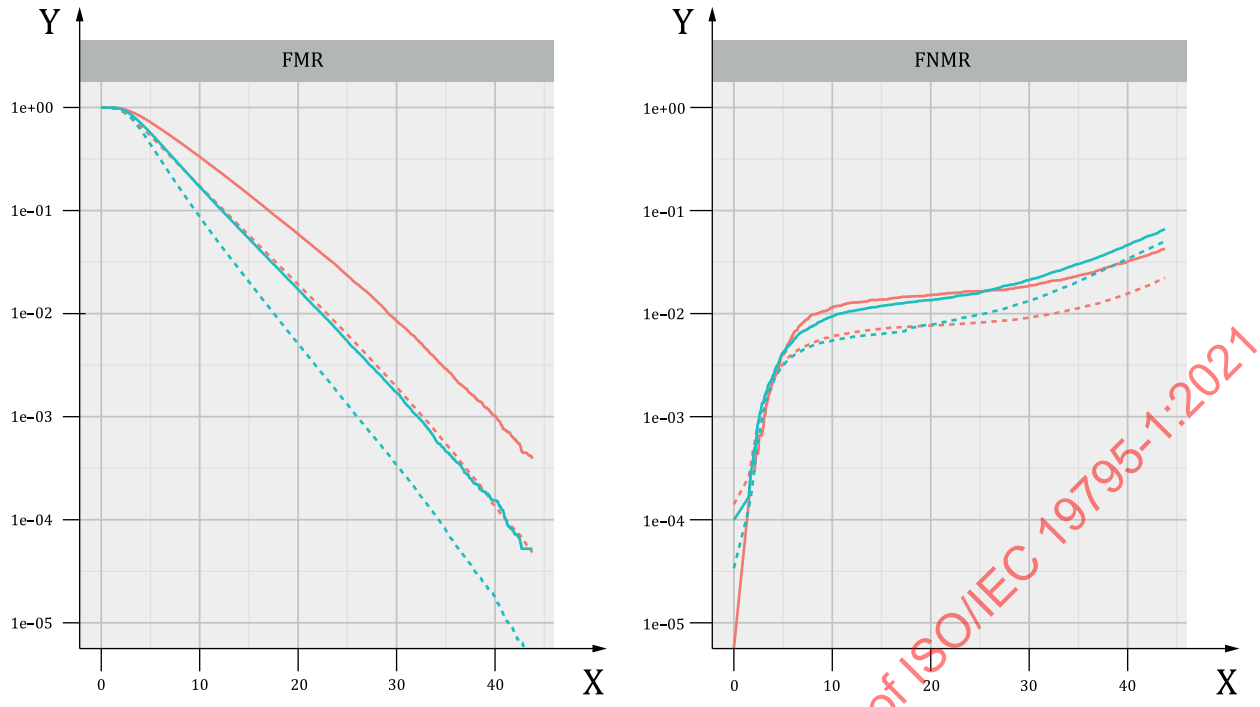
Key

Y comparison score

Figure 3 — Example box and whisker plot

10.2 Error rate vs threshold plot

An error rate may be plotted as a function of a decision threshold, for example as in Figure 4.



Key

X threshold
Y error rate
Race: Black
Race: White
Sex: Female
Sex: Male

Figure 4 — Example: changes in mated and non-mated score distributions due to race and sex

10.3 DET plot

10.3.1 The comparison and/or decision performance of a biometric system over a range of decision thresholds may be graphically represented using a DET plot. DET plots are threshold-independent, allowing performance comparison of different systems under similar conditions, or of a single system under differing conditions.

10.3.2 DET plots may be used to plot decision error rates (false non-match rate against false match rate or false reject rate against false accept rate), or identification error rates (false-negative identification rate against false-positive identification rate). For comparing the performance of different systems, the decision error DET, which shows the combined effect of comparison decision errors, acquisition failures, binning errors, and enrolment failures, can be more helpful than graphs showing the fundamental error rates.

10.3.3 The DET should be plotted with false-positive rate (e.g. FMR, FAR, FPIR) on the abscissa (x-axis) and false-negative rate (e.g. FNMR, FRR, FNIR) on the ordinate (y-axis).

10.3.4 Axis scaling, minimum and maximum values shown, should be selected for clarity of the presented results. To give greater clarity of error rates in the range of interest, DET plots are generally shown using normal deviate, or logarithmic axes.

a) *Linear axis scale:* Plotting the DET with linear axis scaling is not recommended for error rates that span several orders of magnitude. The systems of interest (low error rates) cluster close to the origin, and it is then difficult to distinguish the differences in performance.

b) *Logarithmic axis scale:* Plotting the DET with $\log_{10}$ axis scaling provides more detail at low error rates and helps to distinguish between similarly performing systems.

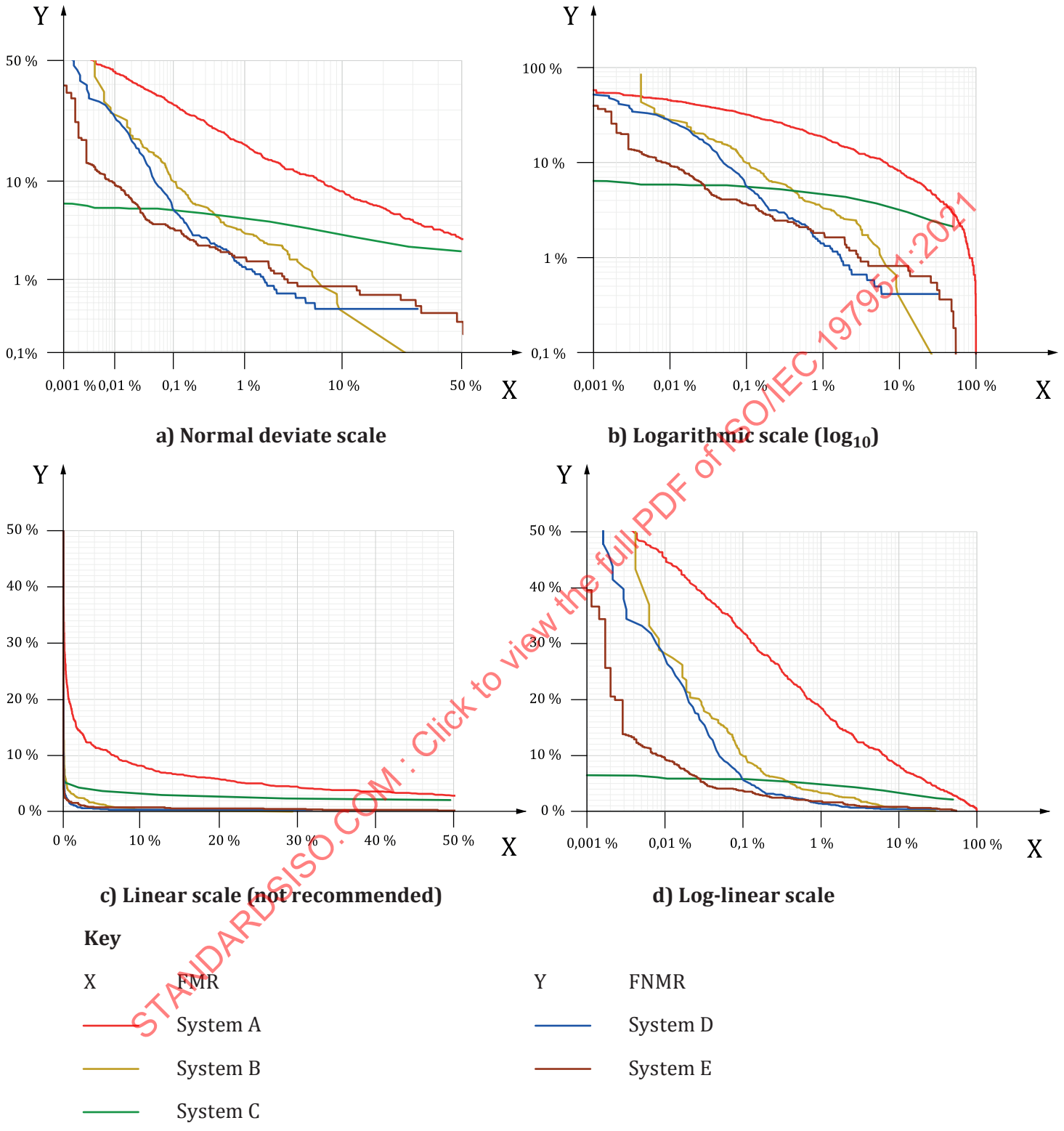
c) *Normal deviate axis scale*: Plotting the DET using normal deviate scaling also provides more detail at low error rates. Often mated and non-mated score distributions are approximately Gaussian, and if so, normal deviate scaling gives DET plots that are roughly linear.

NOTE The DET as originally proposed by Martin et al,[20] was to be plotted using normal deviate scales. In this document other scaling is permitted.

Figure 5 illustrates the effect of different axis scaling for plotting the same DET data and shows that linear axis scaling is inappropriate in this case.

10.3.5 The scaling used shall be reported so that the tabular DET representation can be inferred from the DET plot representation. Axis scaling should be consistent between different graphs in the same report. If it is necessary to change scaling to maintain clarity, there should be a note to the figure remarking on the change of scales.

10.3.6 Figures G.1 to G.4 provide further information to assist readers in interpreting and understanding the information depicted in a DET plot.



a) Normal deviate scale

b) Logarithmic scale ($\log_{10}$)

c) Linear scale (not recommended)

d) Log-linear scale

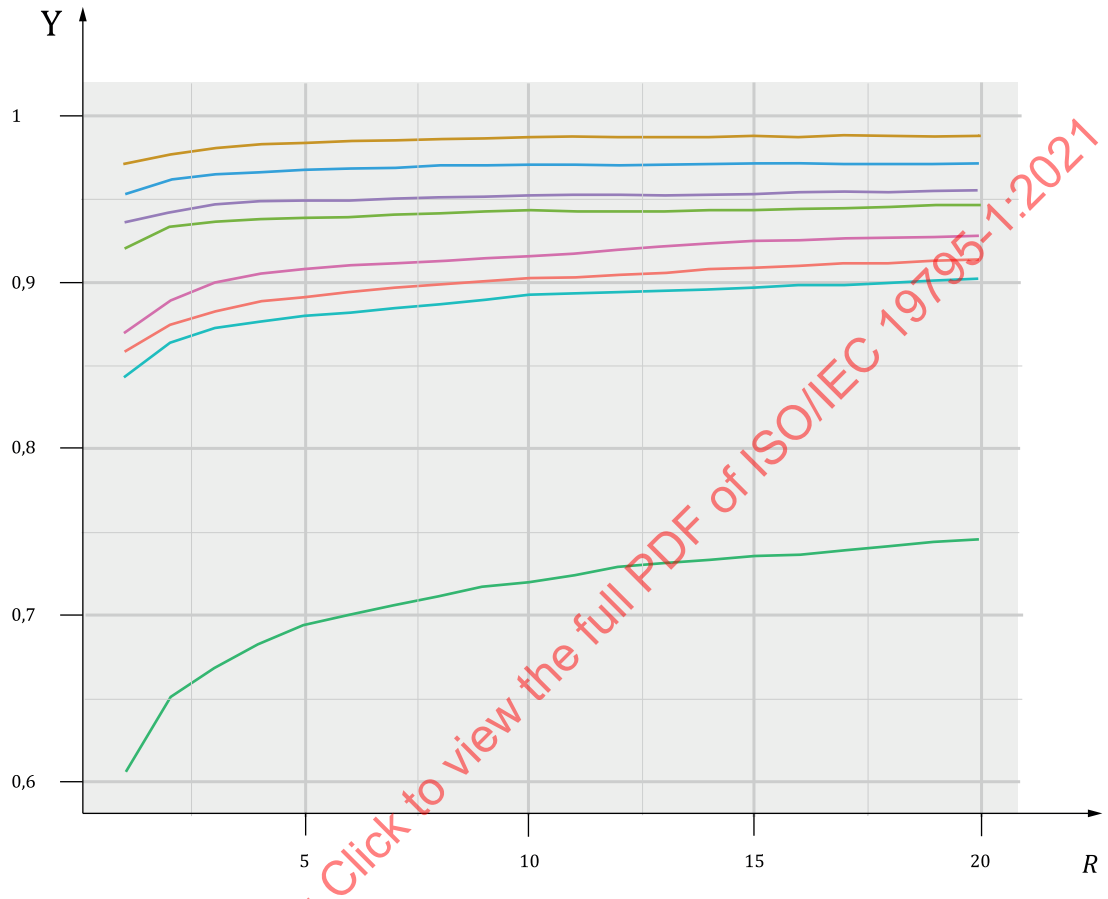
Key

X	FMR	Y	FNMR
	System A		System D
	System B		System E
	System C		

Figure 5 — Example DET plots illustrating effect of axis scaling (a) normal deviate scale, (b) logarithmic scale, (c) linear scale (not recommended), (d) log-linear scales

10.4 CMC plot / FNIR over rank plot

For applications in which the system returns lists of candidates to human operators, performance results are often illustrated using a cumulative match characteristic plot. The curve plots, as a function of R the proportion of transactions where a test subject's identifier is included among the top R identifiers returned. An example is shown in Figure 6.



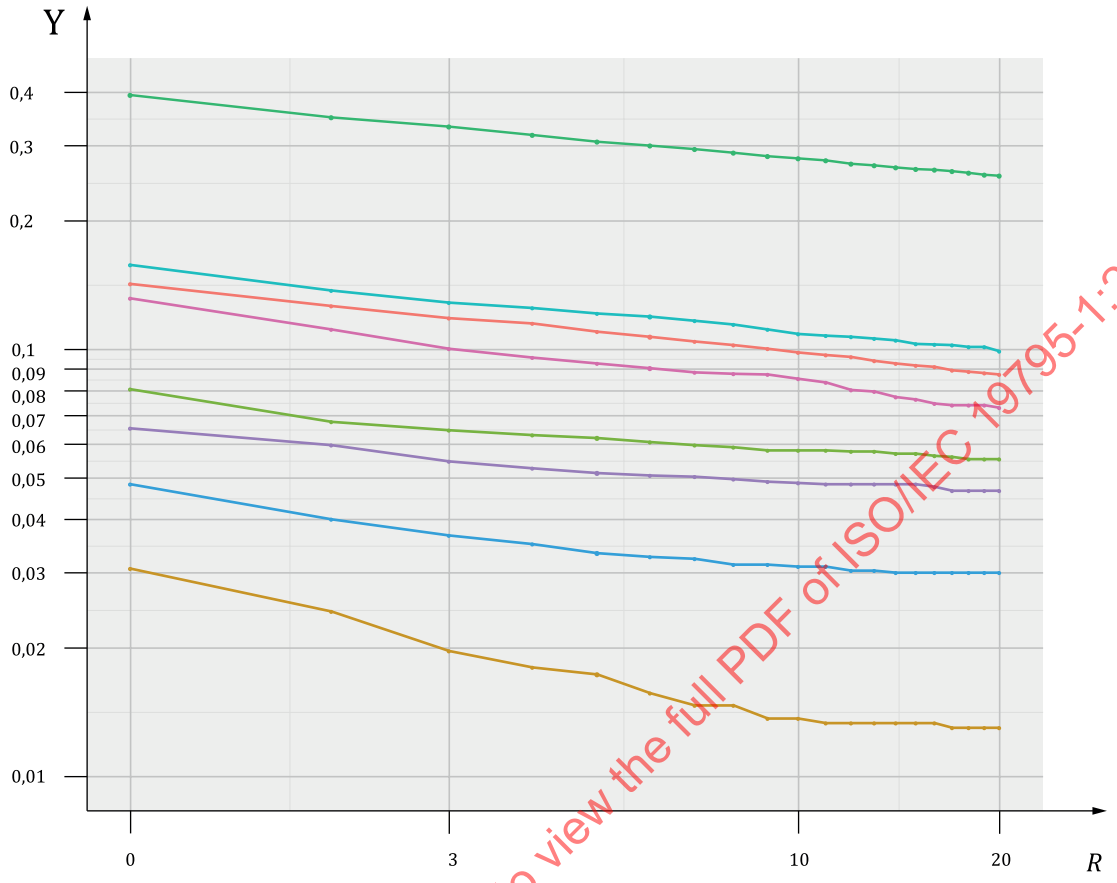
Key

Y true-positive identification rate, TPIR(N=691282, R, T=0) on dataset: Concourse

- Algorithm_C0, TPIR(691281, 1, 0) = 0,97
- Algorithm_F0, TPIR(691281, 1, 0) = 0,95
- Algorithm_F1, TPIR(691281, 1, 0) = 0,93
- Algorithm_C1, TPIR(691281, 1, 0) = 0,92
- Algorithm_G1, TPIR(691281, 1, 0) = 0,87
- Algorithm_A1, TPIR(691281, 1, 0) = 0,86
- Algorithm_E0, TPIR(691281, 1, 0) = 0,84
- Algorithm_D0, TPIR(691281, 1, 0) = 0,61

Figure 6 — Example CMC plot

An alternative, preferred because it directly shows the error rate and not the corresponding true-positive identification rate, is an "FNIR over rank" plot of FNIR(*N*, *R*, 0) against *R*. If the FNIR values span more than one order of magnitude, the FNIR should use a logarithmic scale as shown in Figure 7.



Key

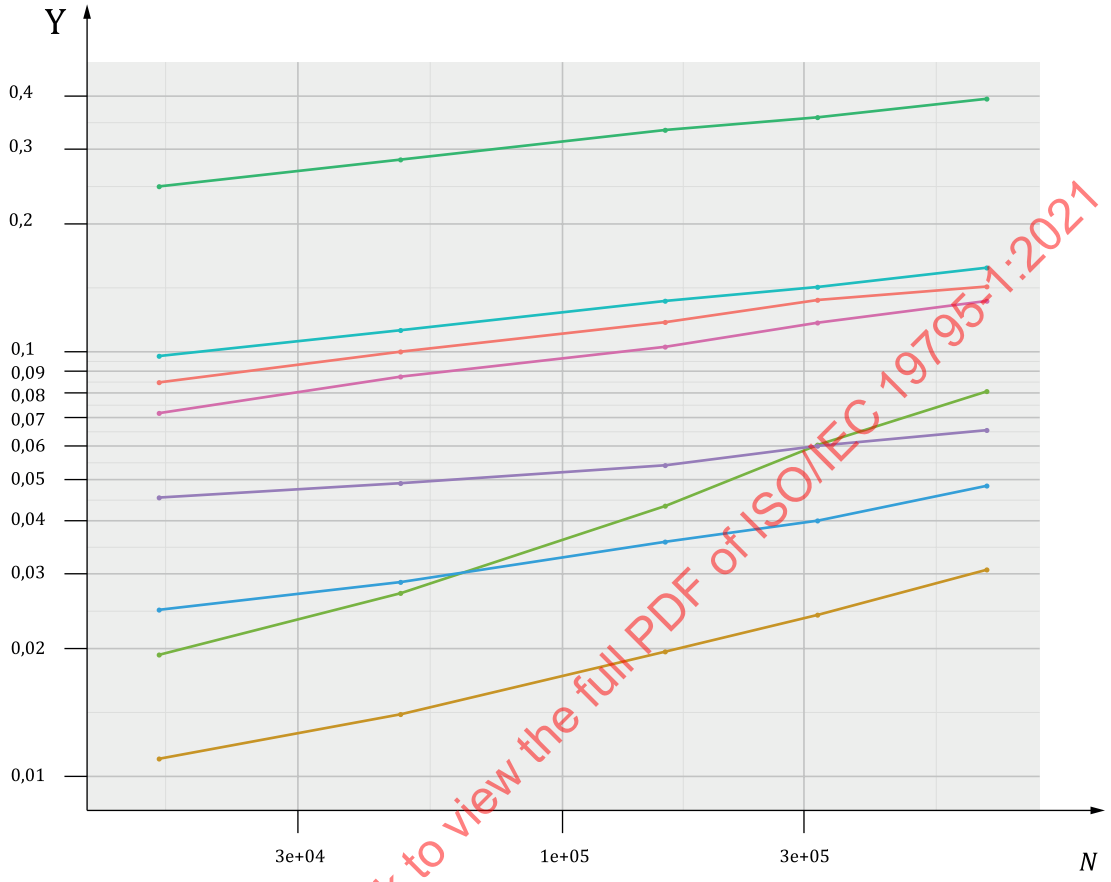
Y false-negative identification rate FNIR(*N*=691282, *R*, *T*=0) on dataset: Concourse

- Algorithm_D0, FNIR(691281, 1, 0) = 0,39
- Algorithm_E0, FNIR(691281, 1, 0) = 0,16
- Algorithm_A1, FNIR(691281, 1, 0) = 0,14
- Algorithm_C1, FNIR(691281, 1, 0) = 0,13
- Algorithm_C1, FNIR(691281, 1, 0) = 0,08
- Algorithm_F1, FNIR(691281, 1, 0) = 0,07
- Algorithm_F0, FNIR(691281, 1, 0) = 0,05
- Algorithm_C0, FNIR(691281, 1, 0) = 0,03

Figure 7 — Example FNIR over rank plot

10.5 FNIR over number of enrolees plot

If searches are performed in enrolment databases of different sizes, performance results should be illustrated using a plot of FNIR(*N*, 1, 0) over number of enrolees *N* as in Figure 8.



Key

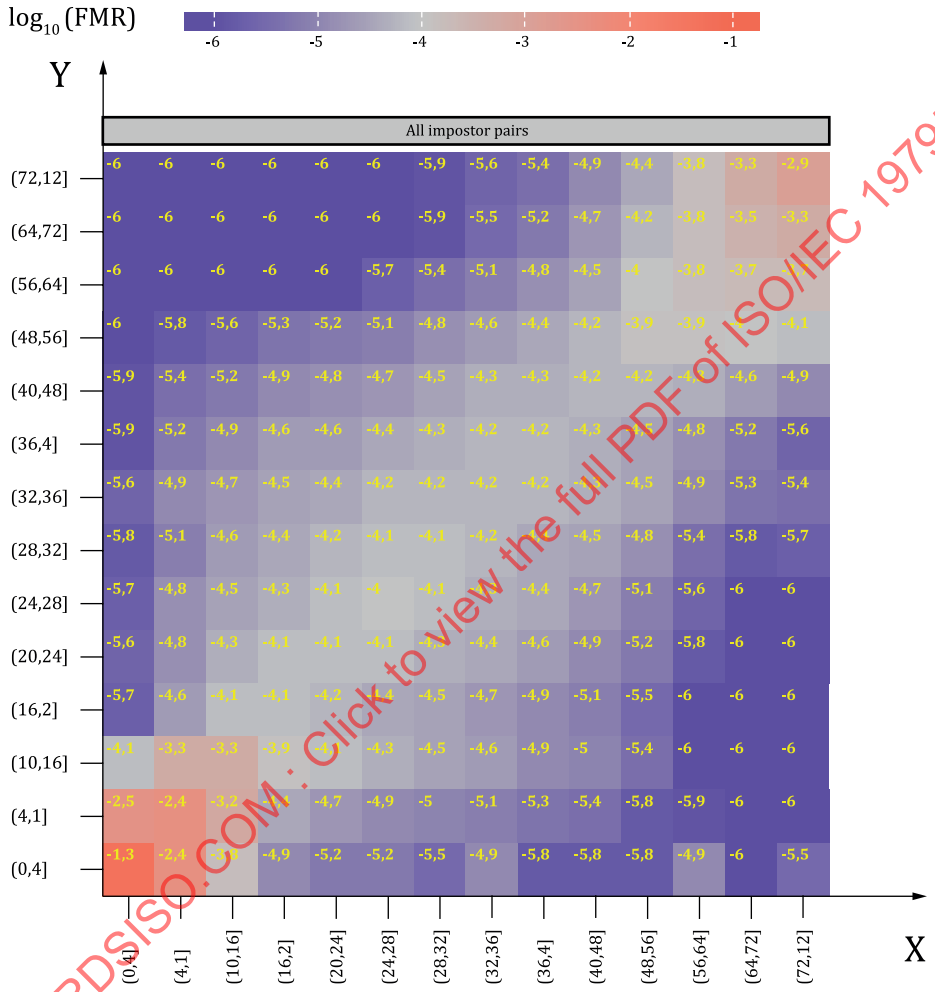
Y false-negative identification rate, FNIR(*N*, *R*=1, *T*=0) on dataset: Concourse

- Algorithm_D0, FNIR(691281, 1, 0) = 0,39
- Algorithm_E0, FNIR(691281, 1, 0) = 0,16
- Algorithm_A1, FNIR(691281, 1, 0) = 0,14
- Algorithm_G1, FNIR(691281, 1, 0) = 0,13
- Algorithm_C1, FNIR(691281, 1, 0) = 0,08
- Algorithm_F1, FNIR(691281, 1, 0) = 0,07
- Algorithm_F0, FNIR(691281, 1, 0) = 0,05
- Algorithm_C0, FNIR(691281, 1, 0) = 0,03

Figure 8 — Example plot of FNIR over number of enrolees

10.6 Heat maps

Heat maps can be useful to illustrate how performance varies over the range of covariates within an experimental design. Figure 9 provides an example heat map showing the variation of false match rate across different ages of impostor and enrolees, when the decision threshold setting was selected to achieve a 1 in 10 000 false match rate over the full set of enrolee-impostor pairs. The colour scale of the heatmap uses a neutral grey at this FMR value, colour shade used tends towards red as the FMR increases above this value, and towards blue as the FMR decreases below this value. The diagonal dominance illustrates that false matches are far more likely to occur when the two images being compared are close in age to one another.



Key

X age of enrolee

Y age of impostor

Figure 9 — Example heat map:
Variation in FMR due to age difference between impostor and enrolee

11 Record keeping

11.1 Record keeping shall comply with regulation concerning retention and storage of personally identifiable information:

a) The original biometric samples, references and features (if collected) shall be stored in conformance with all relevant data protection legislation;

b) Comparison scores, and decisions output by the system, if available, should be retained.

NOTE Comparison scores and decisions are not personally identifiable information unless linked to identities or individualized biometric data.

11.2 Records regarding the methods used to derive performance measures, and the identities of staff responsible for conducting enrolment and supervising the collection of transaction data shall be retained.

11.3 Sufficient information shall be kept to:

a) enable the evaluation to be repeated under conditions as close as possible to the original,

b) facilitate, if possible, identification of factors affecting the uncertainty of results,

c) establish an audit trail, and

d) allow for comparison of results with those of other evaluations, e.g. for inter-laboratory comparison.

11.4 Records (whether written or electronic) shall be protected to avoid loss or change of the original test data. If alterations have to be made, a copy of the original should be kept with a note of the alterations. A mechanism may be required to enable removal of a test subject's personally identifiable information on their request.

11.5 Where mistakes occur (in the data collection procedures etc.) records should show both the original erroneous data, and the corrected values.

11.6 Records shall be disposed of in accordance with the legislation of the relevant jurisdiction and the policy of the organization that has stored the records.

12 Reporting performance results

12.1 Reporting test details

Performance measures such as error rates, transaction times, computational workload, etc. are dependent on test type, application and population. For performance measures to be interpreted correctly, the information in Table 1 shall be reported, together with the performance metrics relevant to the evaluation listed in 12.3 through 12.8.

Table 1 — Reporting test details

Test details	Reporting	Details to report
The system(s) tested	Mandatory	Including details of algorithms, biometric sensors, user interface, supporting hardware, etc.
Test organization details	Mandatory	Test organization, location, date of test.
Type of evaluation	Mandatory	In the case of technology evaluation: details of the test corpus used. In the case of scenario evaluation: details of the test scenario. In the case of operational evaluation: details of the operational application.

Table 1 *(continued)*

Test details	Reporting	Details to report
Size of evaluation	Mandatory	Number of test subjects. Number of instances (fingers, hands or eyes, etc.) enrolled by each test subject. Number of visits made by test subject. Number of transactions per test subject (or test subject instance) at each visit.
Test crew	Mandatory	Demographics of the test crew (age, gender, etc.) The manner in which the test crew was assembled, to include exclusions, volunteers etc., as well as the degree to which the test crew mirrored the target population. The level of training, instruction, familiarization, and habituation of test crew in the use of the system.
Test environment	Mandatory	See 8.3.2.1, 8.4.2, and C.2.6.
Time separation between enrolment and recognition transactions	Mandatory	See 7.3.7.
Quality and decision thresholds used during data collection	Mandatory	The thresholds used, and those recommended for the target application (if different).
Control of factors potentially affecting performance	Mandatory	See 7.3 and Annex C.
Test procedures	Mandatory	E.g. policies for determining enrolment failures. Details of any abnormal cases occurring during testing that are excluded from performance analysis.
Estimated uncertainties	Optional	Estimated uncertainty in performance results, and method of estimation. See 9.11 and Annex B.
Deviation from guidelines	Optional	Deviations from the guidelines of this document should be explained. Sometimes it is necessary to compromise one aspect to achieve another; for example, randomizing the order of using fingers on a fingerprint device might lead to user confusion and a higher number of labelling errors.

12.2 Summary statistics

Single number "summary statistics" such as Equal Error Rate (EER), Half Total Error Rate (HTER), Area under the ROC, are deprecated. If an overall figure of merit is to be provided:

a) the methods to derive such a figure shall be operationally relevant taking account of factors such as error costs, target security level etc. of the system, and

b) the method of derivation shall be reported.

12.3 Reporting enrolment performance

Table 2 lists performance metrics for biometric enrolment.

Table 2 — Enrolment performance metrics

Metric	Reporting	Details to report
Failure to enrol rate (FTER)	Mandatory	See 9.2.1.

Table 2 *(continued)*

Metric	Reporting	Details to report
Enrolment transaction duration	Optional	See 9.2.2. In additional to an average (mean or median) transaction duration, the cumulative distribution function of enrolment transaction times should be provided showing both successful and failed enrolments.
Enrolment transaction computational workload	Optional	See 9.10.2 a).

12.4 Reporting acquisition performance

Table 3 lists performance metrics for the biometric acquisition process.

Table 3 — Acquisition performance metrics

Metric	Reporting	Details to report
Failure-to-acquire rate (FTAR)	Mandatory	See 9.3.1.
Acquisition process duration	Optional	See 9.3.2. In additional to an average (mean or median) transaction duration, the cumulative distribution function of acquisition times should be provided showing both successful and failed acquisitions.

12.5 Reporting one-to-one comparison performance

Table 4 lists performance metrics for one-to-one biometric comparison.

Table 4 — One-to-one comparison performance metrics

Metric	Reporting	Details to report
False match rate (FMR)/ False non-match rate (FNMR)	Mandatory	See 9.4.1 and 9.4.2. FMR and corresponding FNMR shall be reported over the range of decision thresholds tested. A DET plot is recommended in the case of multiple operating points.
FTER	Mandatory	See 12.3. Otherwise a statement that FTER is unknown.
FTAR	Mandatory	See 12.4. Otherwise a statement that FTAR is unknown.
Computational workload of biometric comparison	Optional	

12.6 Reporting verification system performance

Table 5 lists performance metrics for biometric verification.

Table 5 — Verification system performance metrics

Metric	Reporting	Details to report
False accept rate (FAR)/ False reject rate (FRR)	Mandatory	See 9.5.2 and 9.5.3. FAR and corresponding FRR shall be reported over the range of decision thresholds tested. A DET plot is recommended in the case of multiple operating points.
FTER	Mandatory	See 12.3. Otherwise a statement that FTER is unknown.
FTAR	Mandatory	See 12.4. Otherwise a statement that FTAR is unknown.

Table 5 *(continued)*

Metric	Reporting	Details to report
Verification transaction duration	Optional	See 9.5.4. In additional to an average (mean or median) transaction duration, the cumulative distribution function of acquisition times should be provided showing separately accepted and rejected verification durations.
Generalized false accept rate (GFAR)/ Generalized false reject rate (GFRR)	Optional	See 9.5.5. Recommended when comparing systems having different FTER / FTAR error rates. The method of generalization shall be reported.
Verification transaction computational workload	Optional	See 9.10.2 b).

12.7 Reporting identification system performance

Table 6 lists performance metrics for biometric identification.

Table 6 — Identification system performance metrics

Metric	Reporting	Details to report
False-positive identification rate (FPIR)/ False-negative identification rate (FNIR)	Mandatory	See 9.6.2 and 9.6.3. FPIR and corresponding FNIR shall be reported over the range of decision thresholds and identification ranks tested. A DET plot is recommended in the case of multiple operating points. Several DET plots may be shown corresponding to different numbers of identifiers returned, and different number of references in the enrolment database.
Number of enrolled references	Mandatory	
FTER	Mandatory	See 12.3. Otherwise a statement that FTER is unknown.
FTAR	Mandatory	See 12.4. Otherwise a statement that FTAR is unknown.
Selectivity	Optional	See 9.6.5.
Closed-set results	Optional	See 9.6.6. If a closed-set test was conducted, the results should be shown as a CMC plot or as an FNIR-over-rank plot, with details of the number of enrolled subjects.
Identification transaction duration	Optional	See 9.9 In addition to an average (mean or median) transaction duration, the cumulative distribution function of acquisition times should be provided.
Computation workload for an identification transaction	Optional	See 9.10.2 c) and 9.10.4. Computational workload may be measured for different numbers of references to show how workload scales with database size.

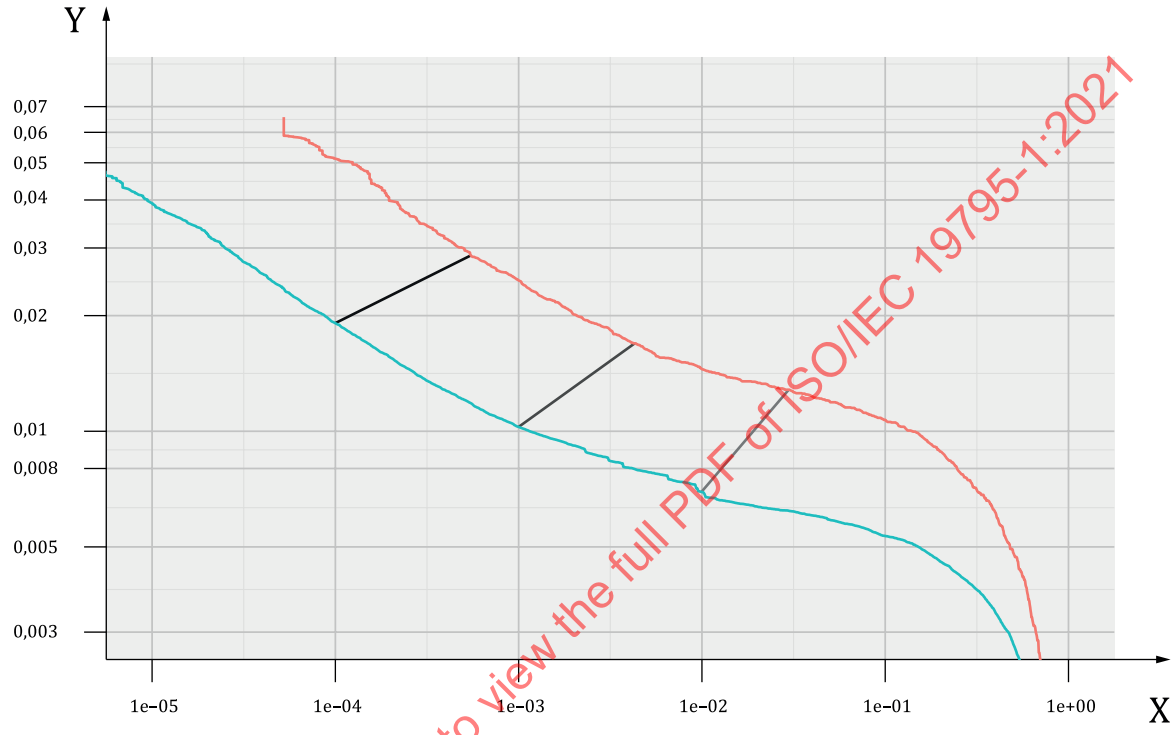
12.8 Reporting performance across factors

12.8.1 It can be useful to show variation in performance between individuals / classes of individuals, for example reporting error rates/comparison scores/timings for different classes of individuals (e.g. male / female). In evaluations where test subjects make multiple mated transactions, it can also be useful to provide histograms showing how individual error rates vary between subjects.

12.8.2 In evaluations considering the effect of a controlled experimental factor on performance, (see 7.3.2a), there can be changes in both FNMR and FMR (or in FNIR and FPIR for an identification system).

The experimenter should produce two graphs, FNMR vs. threshold and FMR vs threshold which include traces for each level of the controlled factor (see examples in Figure 4). Such graphs generally expose the changes in the mated and non-mated distributions respectively.

The normal DET is appropriate to applications where thresholds are set separately for each level of the factor. In cases where a common threshold is used across all levels of the factor, the experimenter may plot DETs, as in Figure 10, annotated by lines connecting points of fixed threshold (shown in black in Figure 10).



Key

X false match rate (FMR)

Y false non match rate (FNMR)

Sex: Female

Sex: Male

Connector between points at same threshold

Figure 10 — Example: Relative performance of male and female subjects for a face-recognition based access control system

Annex A
(informative)

Differences between evaluation types

Table A.1 — Differences between evaluation types

	Technology	Scenario	Operational
What is tested	Biometric component (comparison or extraction algorithm)	Biometric system	Biometric system
Ground truth	Known, subject to data collection errors and intersections in merged data sets	Known, subject to data collection errors and tester failure to note unwanted subject behaviour	Dependent on available controls and instrumentation to establish ground truth
Subject behaviour controlled by test administrator	Not applicable during testing; may be known to be controlled when biometric data recorded, otherwise considered to be uncontrolled	Controlled (unless subject behaviour is an independent variable)	Uncontrolled
Subject has real-time feedback of the result of attempt	No	Yes	Yes
Repeatability of results	Repeatable (corpus fixed)	Quasi-repeatable (if test scenario and population controlled)	Not repeatable
Control of physical environment	May be known to be controlled when biometric data recorded, otherwise considered to be uncontrolled	Controlled and/or recorded	Not controlled, ideally recorded
Subject interaction recorded	Not applicable during testing; may be recorded when biometric data recorded	Recorded	Recorded during enrolment; may be recorded during verification/identification
Typical results reported	Comparison of biometric components or versions of components (e.g. comparison or extraction algorithms or sensors), determine critical performance factors	Compare biometric systems, determine critical performance factors; measure simulated performance	Measure performance in an operational environment
Typical metrics	Most performance metrics (not end-to-end throughput); most error rates; good for large-scale identification system performance where difficult to assemble large test crew	Predicted end-to-end throughput, FMR, FNMR, FTAR, FTER, FAR, FRR	End-to-end throughput; reliable testing of operational FAR and FRR requires some knowledge of ground truth
Constraints	Appropriate test corpus, e.g. gathered with one or more sensors, the identity of which may or may not be known	Operational, instrumented system	Operational, instrumented system; typically only decision rates are available
Human test population	Recorded	Live	Live

NOTE Although in some cases there are exceptions to the entries in this table, these are the mainstream, fundamental characteristics and distinctions.

Annex B
(informative)

Test size and random uncertainty

B.1 Confidence intervals and test size assuming independent identically distributed comparisons

B.1.1 Rule of 3

The Rule of 3 [5],[21]-[23] addresses the issue of the lowest error rate that can be statistically established with a given number, n, of independent identically distributed (i.i.d.) comparisons trials. This value is the error rate p for which the probability of zero errors in n trials, purely by chance, is (for example) 5 %. This gives:

$$p \approx 3/n$$

for a 95 % confidence level.

EXAMPLE A test of 300 independent samples returning no errors is said to have an error rate of 1 % or less with 95 % confidence.

NOTE 1 $p \approx 2/n$ for a 90 % confidence level.

NOTE 2 The i.i.d. assumption is fulfilled, for evaluation of FMR and FNMR, if each mated comparison trial is made by a different test subject, and if each non-mated comparison trial involves a different pair of test subjects. With n test subjects, this would allow only n statistically independent mated comparison trials, and only $n/2$ statistically independent non-mated comparison trials. However, cross-comparisons between all submitted sample features and enrolled references generates many more non-mated comparison trials and, according to Reference [11], achieves smaller uncertainty despite dependencies between the attempts. Thus, except perhaps in the case of operational testing, there is little merit in restricting data to a single attempt per subject to achieve the i.i.d. assumption.

B.1.2 Rule of 30

The Rule of 30 states that to be 90 % confident that the true error rate is within ±30 % of the observed error rate, there should be at least 30 errors [24]. So, for example, if there are 30 false non-match errors in 3 000 independent mated comparison trials, we can say with 90 % confidence that the true error rate is between 0,7 % and 1,3 %. The rule comes directly from the binomial distribution, assuming independent trials, and may be applied by considering the performance expectations for the evaluation.

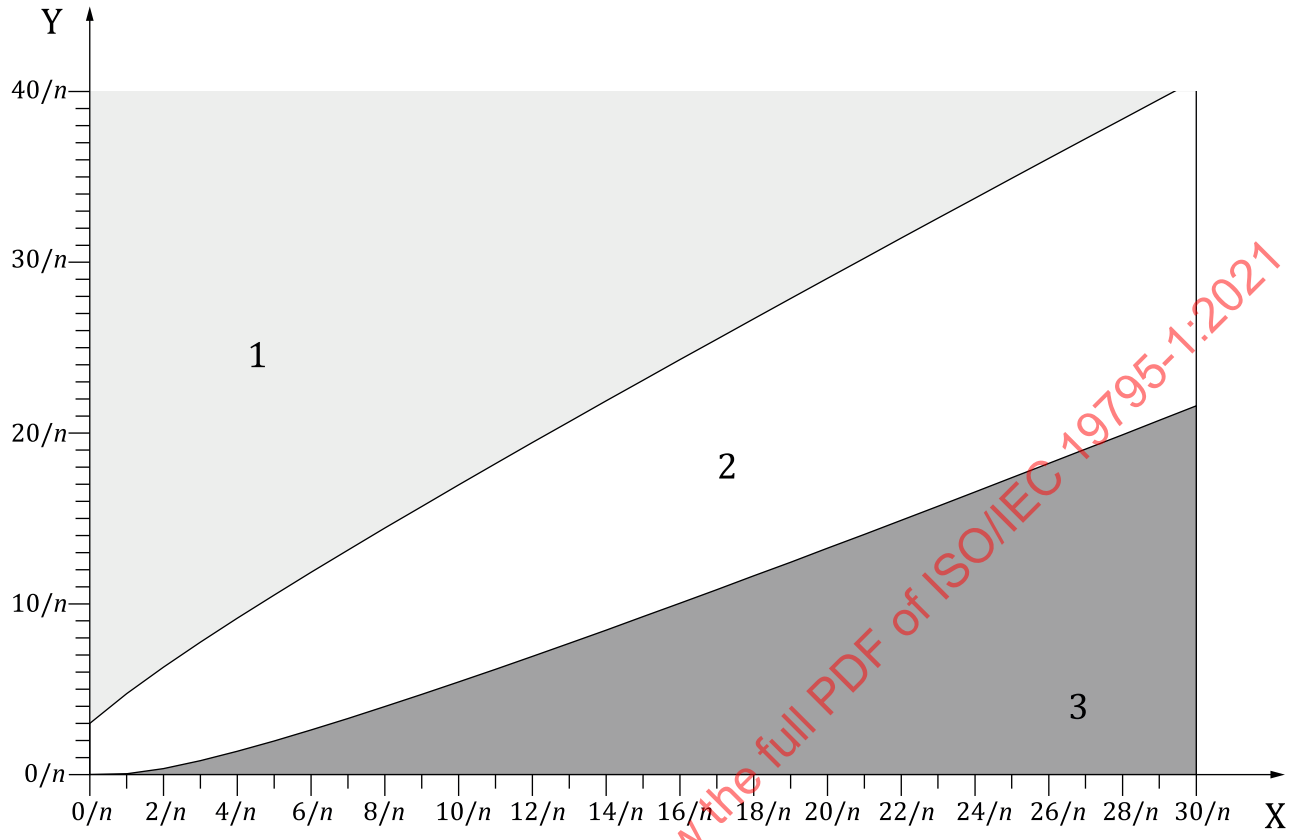
EXAMPLE Suppose the performance goals are a 1 % false non-match rate, and a 0,1 % false match rate. This rule implies 3 000 mated comparison trials and 30 000 non-mated comparison trials. Note however, to fulfil the requirement of independence of comparison trials, would require 60 000 test subjects. The alternative is to compromise on independence by re-using a smaller set of test subjects, and to be prepared for a loss of statistical significance.

NOTE The rule generalizes to different proportional error bands. For example, to be 90 % confident that the true error rate is within ±10 % of the observed value, at least 260 errors are needed. To be 90 % confident that the true error rate is within ±50 % of the observed value, at least 11 errors are needed.

B.1.3 Number of comparisons to support a claimed error rate

B.1.3.1 The number of statistically independent comparisons required to support a claimed error rate is illustrated in Figure B.1. For example, no false matches in n independent non-mated comparisons

would support a claimed false match rate of $3/n$, with 95 % confidence, while 30 errors would support a claim of $41/n$.



Key

X errors observed in *n* independent comparisons
Y errors rate claimed
1 claim supported *(95 % confidence)*
2 claim neither supported nor refuted
3 claim refuted *(95 % confidence)*

NOTE This chart provides a reasonable approximation when the claimed error rate is 1 % or below.

Figure B.1 — 95 % confidence decision regions for accepting (or rejecting) an error rate claim with *n* independent comparisons

B.1.3.2 To ensure statistical independence, the test subjects providing the probes and references in each non-mated comparison trial need to be different and selected randomly from the target population. This approach is unlikely to be efficient for low false match rates, as *n* independent comparisons requires $2n$ test subjects.

B.1.3.3 An alternative cross-comparison approach is often adopted, though this does not ensure statistical independence. With *n* individuals, cross-comparison of attempts/references for each (unordered) pair, can exhibit a low degree of correlation. The correlations within these $n(n-1)/2$ false match attempts reduce the confidence level for supporting an FMR claim compared with the same number of completely independent comparisons.

B.2 Variance of performance measures as a function of test size

As the test size increases, the variance of estimates decreases, but the scaling factor depends on the source of variability.

If test subjects each make multiple mated comparison trials, then the variance of the observed false non-match rate has components due to:

— variability of test subjects, scaling as 1 / (number of test subjects); and

— residual variability of mated comparison trials, scaling as 1 / (number of mated comparisons).

If test subjects make multiple attempts, and non-mated comparison trials are generated offline by cross-comparison of these attempts against references from a different set of data subjects, then the variance of the observed false match rate has components due to:

— variability of test subjects, scaling as 1 / (number of test subjects);

— variability of impersonated references, scaling as 1 / (number of impersonated references);

— variability of attempts (other than that accounted for by variability of test subjects), scaling as 1 / (number of attempts); and

— residual variability of the generated non-mated comparison trials, scaling as 1 / (number of non-mated comparison trials).

NOTE Doddington et al.[3] show that biometric systems can have "goats", "lambs" and "wolves". Goats have a personal false non-match rate significantly higher than that for the overall population, lambs are those whose references incur a disproportionate share of false matches, while wolves are those whose samples are particularly successful at giving false matches. This would imply that, for the false non-match rate, the component of variance for test subjects is non-zero; and for the false match rate, the components for test subjects and for references are non-zero.

B.3 Estimates for variance of performance measures

B.3.1 General

This subclause presents formulae and methods for estimating the variance of performance measures. The variance is a statistical measure of uncertainty and can be used in estimating confidence intervals, etc. The applicability of these formulae depends on the following assumptions about the distribution of error cases:

a) The test crew is representative of the target population. This is be the case if, for example, the test subjects are drawn at random from the target population.

b) Attempts by different test subjects are independent. This is not always be true. Subject interactions are influenced by what they see others do. However, the correlations between test subjects are likely to be minor in comparison to the correlations within a set of attempts by one test subject.

c) Attempts are independent of threshold. Otherwise, the estimates for the error rates can be biased except at the threshold used for data collection.

d) Error rates vary across the population. Different subjects can have different individual false non-match rates, and different subject pairs can have different individual false match rates.

e) The number of observed errors is not too small. In cases with no observed errors, the formulae would give a zero variance, but the Rule of 3 would apply. Schuckers et al.[25] examines the conditions under which the methods such as those presented here provide appropriate coverage intervals.

B.3.2 Variance of observed false non-match rate

B.3.2.1 False non-match rate — Single attempt per test subject

In the case where each test subject makes a single attempt, Formulae (B.1) and (B.2) apply:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} a_i \tag{B.1}$$

$$\hat{v}\left(\hat{p}\right) = \frac{\hat{p}\left(1-\hat{p}\right)}{n-1} \tag{B.2}$$

where

n is the number of enrolled test subjects;

a_i is the number of false non-matches for the i^{th} test subject;

$\hat{p}$ is the observed false non-match rate;

$\hat{v}\left(\hat{p}\right)$ is the estimated variance of observed false non-match rate.

NOTE 1 A derivation of this estimate can be found in many statistical textbooks (e.g, Reference [26]).

NOTE 2 These formulae have sometimes been misapplied to cases where subjects make several attempts. The replacement of the number of test subjects, n, by the number of attempts is generally not valid.

NOTE 3 These formulae are also appropriate for estimating variances of failure-to-acquire and failure-to-enrol rates when there is one attempt per test subject.

B.3.2.2 False non-match rate — Multiple attempts per test subject

In the case where each test subject makes the same number of attempts, the appropriate estimates are given by Formulae (B.3) and (B.4)[26]:

$$\hat{p} = \frac{1}{mn}\sum_{i=1}^{n} a_i \tag{B.3}$$

$$\hat{v}\left(\hat{p}\right) = \frac{1}{(n-1)}\left(\frac{1}{m^2 n}\sum_{i=1}^{n} a_i^2 - \hat{p}^2\right) \tag{B.4}$$

where:

n is the number of enrolled test subjects;

m is the number attempts made by each test subject;

a_i is the number of false non-matches for the i^{th} test subject;

$\hat{p}$ is the observed false non-match rate;

$\hat{v}\left(\hat{p}\right)$ is the estimated variance of observed false non-match rate.

NOTE 1 When $m = 1$, the estimates are the same as those in Formulae (B.1) and (B.2).

NOTE 2 These formulae are also appropriate for estimating variances of failure-to-acquire rates when there are multiple attempts per test subject.

B.3.2.3 False non-match rate — Unequal numbers of attempts per test subject

Sometimes the number of attempts per subject varies. Some subjects might not complete the desired number of attempts. Acquisition process failures can also cause attempts to be missing from the false non-match rate calculations. Provided there is no correlation between the number of attempts made and the differing success rates of individuals, Formulae (B.5) and (B.6) are appropriate:

$$\hat{p} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} m_i} \tag{B.5}$$

$$\hat{v}\left(\hat{p}\right) = \frac{\sum_{i=1}^{n} a_i^2 - 2\hat{p}\sum_{i=1}^{n} a_i m_i + \hat{p}^2 \sum_{i=1}^{n} m_i^2}{\frac{n-1}{n}\left(\sum_{i=1}^{n} m_i\right)^2} \tag{B.6}$$

where:

n is the number of enrolled test subjects;

m_i is the number attempts made by the i^{th} test subject;

a_i is the number of false non-matches for the i^{th} test subject;

$\hat{p}$ is the observed false non-match rate;

$\hat{v}\left(\hat{p}\right)$ is the estimated variance of observed false non-match rate.

NOTE 1 This formula for the variance (from Reference [26]) is an approximation to give an expression in a usable form.

NOTE 2 When all m_i are equal, the same estimates as in Formulae (B.3) and (B.4) are obtained.

NOTE 3 Sometimes the different frequency of use by test subjects is correlated with the differing success rates. For example, test subjects who are rejected make additional attempts to be recognized, while those using the system more frequently achieve better performance through the effects of habituation. In such cases, Formulae (B.5) and (B.6) cannot be directly applied, as there is a risk that results are dominated by a small group of excessively frequent but unrepresentative test subjects.

B.3.3 Variance of observed false match rate

In the case where a full set of cross-comparisons is made, the observed false match rate, and an estimate of the variance are given by Formulae (B.7) and (B.8):

$$\hat{q} = \frac{1}{mn(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} \tag{B.7}$$

$$\hat{v}\left(\hat{q}\right) \approx \frac{1}{m^2 n^2 (n-1)^2} \sum_{i=1}^{n} \left(c_i + d_i\right)^2 - \frac{4}{n}\hat{q}^2 \tag{B.8}$$

where:

n is the number of test subjects (and of enrolment references);

m is the number of samples per test subject;

b_{ij} is the number of samples from the i^{th} test subject falsely matching the reference of the j^{th} test subject (and $b_{ii} = 0$);

c_i is the number of false matches in total against the reference of the i^{th} test subject ($c_i = \sum_{j=1}^{n} b_{ji}$);

d_i is the number of false matches in total by the i^{th} test subject ($d_i = \sum_{j=1}^{n} b_{ji}$);

$\hat{q}$ is the observed false match rate;

$\hat{v}(\hat{q})$ is the estimated variance of the observed false match rate.

For a more general method, not requiring exactly m samples per subject, and allowing for non-square matrices of reference and probe numbers see also Computational Methods in Biometric Authentication, Chapter 4 Equations 4.1 through 4.12[27].

B.4 Estimating confidence intervals

B.4.1 General

B.4.1.1 With a sufficiently large number of attempts, the central limit theorem[26] implies that the observed error rates are expected to follow an approximately normal distribution. However, because we are dealing with proportions near to 0 %, and the variance in the measures is not uniform over the population, some skewness is likely to remain until the number of test subjects is quite large.

B.4.1.2 Under the assumption of normality, 100(1 – α) % confidence bounds on the observed error rates are given by Formula (B.9):

$$\hat{p} \pm z(1-\alpha/2)\sqrt{\hat{v}(\hat{p})} \tag{B.9}$$

where:

$z(\,)$ is the inverse of the standard normal cumulative distribution, i.e. the area under the standard normal curve with mean 0, variance 1 from $-\infty$ to z(x) is x. For 95 % confidence limits, the value of z(0,975) is 1,96;

α is the probability that the confidence interval does not contain the true value of the error rate;

$\hat{p}$ is the observed error rate;

$\hat{v}(\hat{p})$ is the estimated variance of the error rate.

B.4.1.3 Often when the above formula is applied, the confidence interval reaches into negative values for the observed error rate, but negative error rates are impossible. This is due to non-normality of the

distribution of observed error rates. Non-parametric methods, such as the bootstrap can be used to obtain confidence intervals in such cases[28]-[30].

B.4.2 Bootstrap estimates of the variance and confidence intervals

B.4.2.1 Bootstrap estimation reduces the need to make assumptions about the underlying distribution of the observed error rates and the dependencies between attempts. The distributions and dependencies are inferred from the data itself. By sampling with replacement from the original data, a bootstrap sample can be created, from which an alternative estimate of the error rate would be produced. With a large number of such bootstrap samples, an empirical distribution for the estimators can be obtained. This can be used to construct confidence intervals, estimate uncertainties, etc.

B.4.2.2 To illustrate the process, suppose we are estimating the false match rate using a full set of cross comparison with n test subjects, each providing m attempts to be compared against all $(n - 1)$ non-self references. If $x(v, a, t)$ denotes the result of the matching of the a^{th} attempt by test subject v against reference t. The dataset X for estimating the false match rate consists of the results of all $mn(n - 1)$ cross-comparisons $X = \{ x(v, a, t) \mid t \neq v \in \{1, ..., n\}, a \in \{1, ..., m\} \}$. Each bootstrap sample shall be constructed from X in a way that replicates the structure and dependencies in the original data. The procedure is as follows:

a) sample n test subjects with replacement: $v(1), ..., v(n)$. (Sampling with replacement means the list is likely to contain more than one occurrence of the same item);

b) for each $v(i)$ sample with replacement $(n - 1)$ non-self references: $t(i, 1), ..., t(i, n–1)$;

c) for each $v(i)$ sample with replacement m attempts made by that test subject: $a(i, 1), ..., a(i, m)$;

d) the bootstrap sample produced is:

$Y = \{ (v(i), t(i, j), a(i, k)) \mid i \in \{1, ..., n\}, j \in \{1, ..., n–1\}\ a \in \{1, ..., m\} \}$.

Many bootstrap samples are generated, and a false match rate obtained for each. The distribution of the bootstrap values for the false match rate is used to approximate that of the observed false match rate.

B.4.2.3 The bootstrap values allow a direct approach for constructing $100(1 - \alpha)$ % confidence limits: choosing L (lower limit) and U (upper limit) such that only a fraction $\alpha/2$ of bootstrap values are lower than L, and $\alpha/2$ bootstrap values are higher than U. At least 1 000 bootstrap samples should be used for 95 % limits, and at least 5 000 bootstrap samples for 99 % limits.

B.4.3 Subset sampling

B.4.3.1 A further approach to inferring the error margin on the observed error rates is to divide the collected test data into disjoint subsets of test subjects, and then generating a DET for each subset. The FRVT2002 evaluation[31] for example, used this approach to generate error ellipses.

B.4.3.2 The basic approach to deriving error ellipses is as follows.

a) Gather performance results using n test subjects.

b) Divide test population into m (e.g. m =10) disjoint sets of size n/m.

c) Compute DET for each subset.

d) Assume a threshold t.

   1) Find $\boldsymbol{x}_i = (FMR_i, FNMR_i)$ at the threshold for all sets $i = 1, ..., m$.

   2) Compute the sample mean, $\bar{\boldsymbol{x}}$, and sample covariance, S, using Formulae (B.10) and (B.11):

$$\bar{\boldsymbol{x}} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i \quad \text{(B.10)}$$

$$S = \frac{1}{m-1}\sum_{i=1}^{m}\left((\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\right) \quad \text{(B.11)}$$

3) Compute the eigenvectors and eigenvalues of S.

4) The ellipse centroid is at $\bar{\boldsymbol{x}}$, the axis orientations are given by the eigenvectors of S, and the semi-axis lengths are proportional to the square root of the corresponding eigenvectors, the constant of proportionality being the square root of the chi-square factor for the desired confidence level (at 2 degrees of freedom) divided by $\sqrt{m}$.

5) Under the assumption of normality, the error ellipse provides a confidence bound on the values for FMR and FNMR at threshold t (calculated over the whole test population).

e) Repeat for further thresholds t.

NOTE An analogous procedure can be used for identification systems by splitting the search set. For estimation of uncertainty in FPIR, the non-mated search set is split into M disjoint sets. For estimation of certainty in FNIR, the mated search set is split.